Understanding Community Structure for Large Networks

by
Beate Franke

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Supervisor: Patrick J. Wolfe
Department of Statistical Science
University College London
September 30, 2016

I, Beate Franke, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

September 30, 2016

(Beate Franke)

Abstract

The general theme of this thesis is to improve our understanding of community structure for large networks. A scientific challenge across fields (e.g., neuroscience, genetics, and social science) is to understand what drives the interactions between nodes in a network. One of the fundamental concepts in this context is community structure: the tendency of nodes to connect based on similar characteristics.

Network models where a single parameter per node governs the propensity of connection are popular in practice. They frequently arise as null models that indicate a lack of community structure, since they cannot readily describe networks whose aggregate links behave in a block-like manner. We generalize such a model called the degree-based model to a flexible, nonparametric class of network models, covering weighted, multi-edge, and power-law networks, and provide limit theorems that describe their asymptotic properties.

We establish a theoretical foundation for modularity: a well-known measure for the strength of community structure and derive its asymptotic properties under the assumption of a lack of community structure (formalized by the class of degree-based models described above). This enables us to assess how informative covariates are for the network interactions. Modularity is intuitive and practically effective but until now has lacked a sound theoretical basis. We derive modularity from first principles, and give it a formal statistical interpretation. Moreover, by acknowledging that different community assignments may explain different aspects of a network's observed structure, we extend the applicability of modularity beyond its typical use to find a single "best" community assignment.

We develop from our theoretical results a methodology to quantify network community structure. After validating it using several benchmark examples, we investigate a multi-edge network of corporate email interactions. Here, we demonstrate that our method can identify those covariates that are informative and therefore improves our understanding of the network.

Acknowledgments

Four great years go to an end of intense discussions, hard work, sleepless nights and many incredibly rewarding moments when I finally truly understood.

I would first like to thank my supervisor Professor Patrick Wolfe for his support, guidance, and for hours of insightful discussions. Patrick's passion for mathematics and rigor has often motivated me. I strongly appreciate the many doors Patrick has opened for me, and all the exceptional experiences that came with it. I am grateful to Patrick for teaching me to believe in myself, to live up to my potential and for encouraging me to reach even higher.

I would also like to thank the Stochastic Processes Group at UCL, for their curiosity and the great seminars with many enlightening questions. A special mention is needed for Professor Sofia Olhede, Dr Pierre-André Maugis, and Dr Simon Lunagomez. To Sofia for her great advice, creativity and energy. To Pierre-André and Simon for many inspiring discussions, for sharing their experience, and always readily lending an ear.

I am sincerely grateful to everyone at the UCL Department of Statistical Science for creating an open, and stimulating environment, where you can always ask and where people are happy to help. I enjoyed our lunchtimes, chats over tea and Friday nights at ULU. I particularly would like to thank Dr Codina Cotar, Professor Tom Fearn, and Dr Ioanna Manolopoulou for career advice and general support; as well as Dr Ioannis Kosmidis, and Dr Yvo Pokern for inspiring conversations. Special thanks go to Anne-Marie, Rodrigo, Anna, Katrin, Hannah and Verena for moral support; and to Bibi, Mike, Michael, James, Rui, Bryan, Francesco, and Sam for making sure I see more than my books during my time at UCL.

I also greatly acknowledge the funding of my PhD studentship by the UCL Department of Statistical Science. I thank Dr Leon Danon for sharing the data on jazz musicians from [59] and María Dolores Alfaro Cuevas for producing Figure 4.1.

I am thankful to Professor Nancy Reid for inviting me to participate for three months in the Fields program on big data. The people I met there and the talks and discussions I attended, widened my professional horizon. Special thanks go here to Dr Jean-François Plante and Dr Ribana Roscher. On a similar note, I wish to thank Dr Aaron Clauset, Professor Mason Porter

and Dr David Kempe for organizing the Mathematical Research Community on Networks; in particular Mason for career advice and general support. I thank Dr Bailey Fosdick and Professor Gesine Reinert for the interesting discussions.

I also wish to thank Professor Iris Pigeot, and Dr Ronja Foraita who guided me during the very early stages of my career. It is due to Professor Pigeot's passion for statistics and the joint work with Ronja that I found my way to statistics.

I am very grateful to have met Tianmiao and Lisanne, who became very dear to me within the past four years. Thanks for sharing your thoughts and challenges, and making me take a break. A problem shared is a problem halved.

Thanks to my beloved friends Anki, Caro, Janne and Mary for sharing the good and the bad times with me, for always being understanding, and visiting me across the miles—with or without my consent. For our friendship, it seems that time and space do not matter.

I am eternally indebted to my family Marco, Na, Max Mustermann, Moni, Andi, Thore, Geli, Henna, Melanie, Guido, Marlene, Caspar, Marianne, Helmut and Ute for always supporting me in whatever I want to do; for never missing an opportunity to cheer me up and for always making me feel loved and welcome at home. Special thanks to Marlene, Caspar and Thore for reminding me that there are more important things in life.

Finally, I am deeply thankful to my partner Matthias for encouraging me to find my own way, and be myself. I am grateful for your confidence, peace of mind, and attitude that there is no mountain too high. Thanks for being there for me, for believing in me and for making me laugh, even in moments when it seems impossible. I love you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In Chapters 1 and 2, we first motivate the problem addressed in this thesis and then review the literature starting with an introduction of networks, their properties and models; and finishing with a detailed discussion of the prominent challenges in network modeling. Having established a framework and given the right context, we then in Chapters 3–6 turn to our original contributions.

1.1 Motivation

1.1.1 Networks and community structure

In many sciences there has been a conceptual shift away from the study of individual entities and towards the analysis of entire systems—not least because of the technological advances that enable us to collect the corresponding data [79]. In every system, these entities interact either directly or induced as a summary of their dependencies. Networks give us a means to describe and analyze these interactions between entities. In contrast to classical statistics, networks allow us to model complex dependencies while assuming very little structure. For instance, there is no natural ordering and thus no geometry inherited in a network as it is in time series or spatial statistics.

The structure of many networks is strongly influenced by a natural division into communities: sets of nodes with a stronger tendency to connect with nodes of the same set than with nodes of other sets. These communities are often implied by shared characteristics; but may also result from a similar function within the network. Much work has focused on identifying the single "best" community structure. However, one knows that clustering algorithms always return clusters, even when the input is purely noise. Hence, these "optimal" community assignments lack interpretability and present a barrier to understanding.

Understanding the community structure of large networks is crucial to enable statistical modeling and inference on networks that come with provable guarantees. Scientists inevitably observe not only the network nodes and their connections, but also additional information in the form of covariates. By acknowledging that each covariate may explain different aspects of a network's structure, we extend the concept of communities beyond its typical use in the search for a single "best" community assignment. We use covariates to define community assignments, and then deliver a method to quantify how well these communities explain the network's structure.

1.1.2 Motivating examples of networks in science

Because of the ubiquity of networks, a contribution to network analysis has the potential to influence a wide variety of sciences. We now illustrate the importance of networks and community structure on examples in neuroscience, genetics and social sciences.

In *neuroscience*, particularly functional magnetic resonance imaging (fMRI), a volume containing the subject's brain is discretized into three-dimensional voxels whose intensities are measured across time as an index for neural activity. Using the activity to define a similarity measure between voxels, fMRI images of the human brain are often modeled as networks [18, 36, 93]. Because of the size of fMRI data, and since they are naturally affected by structured spatial correlations and high-frequency noise, it is a common approach to combine voxels into functionally distinct regions of interest. This clustering of the network allows us to marginalize the structured short-range dependencies and to reduce the dimensionality [36]. Building up on this, Martino et al. [95] analyze the voxel intensities for two of the communities in patients suffering from bipolar disorder and show in a case-control study with 100 patients that the ratio of intensities is a potential biomarker to distinguish between depressive and mania. Ramot et al. [121] and others [130, 132] go one step further by conducting an intervention study on 16 patients, demonstrating that we may train patients (with and without their awareness) to change the functionality of the communities in the cortical network of the human brain.

In *genetics*, we identify proteins that are strongly associated with a disease and then design drugs such that they modulate these proteins to perturb a disease state. Recent advances indicate that many effective drugs modulate multiple targets instead of a single protein [70, 122]. In this context, modeling protein interactions as a network enables us to analyze the consequences of a drug on the entire system: on the targeted protein; on other proteins that might influence the same phenotype; as well as on off-target proteins that lead to side-effects.

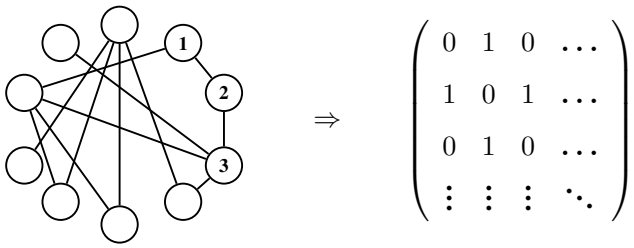
As a consequence, Hopkins and others [70, 122, 155] identify understanding the underlying protein interaction network as one of the main challenges of drug discovery. Vinayagam et al. [141] cluster the proteins into "indispensable", "neutral", or "dispensable" based on their centrality in the network. Based on a study of 1,547 cancer patients, the authors show that disease-causing mutations and drugs target primarily the indispensable proteins and identify 56 genes to be associated with cancer, of which 46 have previously not been known. In contrast, Wu et al. [148] cluster genes into overlapping communities corresponding to regions involved in coherent developmental programs. The authors demonstrate on a study of 1,640 images of the gene expression data of Drosophila (fruit fly) embryos that the communities identified using their unsupervised learning algorithm agree with the well-studied gap gene network.

Social networks of people and their interactions have probably the highest public awareness, not least due to popular examples such as Facebook and LinkedIn. They also have one of the longest academic histories with scientific contributions dating back at least to the 1930's [101]. Of particular importance in social networks is homophily: the tendency of nodes to connect in communities of similar nodes [98]. Kearns et al. [76] exploit homophily to address privacy issues in a clustering task in the context of counterterrorism. The authors cluster people in a social network into target and non-target communities based on their connections. While protecting the privacy rights of non-target individuals, the authors minimize the number of tests; e.g., surveillance, needed to clarify a node's affiliation for certain. Paluck et al. [118] conduct an intervention study to promote anti-conflict behavior in a social network of 56 schools with 24,191 students. While randomization is crucial to the result, the authors first block for predefined community structures; e.g., gender and grade, before randomly selecting schools and students for the intervention. The authors thereby adjust for homophily. Comparing intervention schools against the controls, the disciplinary reports of student conflicts reduced by 30% over 1 year. Thus, the authors demonstrate that influencing a few individuals in a network may be sufficient to cause a behavioral change in the majority of the nodes, when adjusting for homophily.

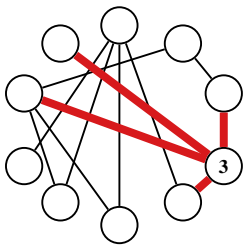
1.2 Preliminaries

1.2.1 Definitions

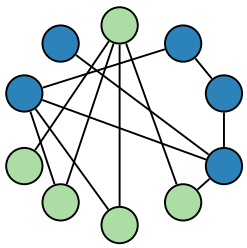
We define a *graph* (i.e., network) $G = (V, E)$ as a tuple of the set of *nodes* V and the set of *edges* $E \subseteq V \times V$. We call the number of nodes $n = |V|$ the *size* of the graph and the *density*



(a) Each network may be represented by an adjacency matrix.



(b) The degree of a node is the sum of its connections ($d_3 = 4$).



(c) Community assignments are partitions of the nodes.

Figure 1.1: Toy example of a binary network with 10 nodes.

denotes the proportion of observed edges N over the number of possible edges:

$$\text{den}(G) = N \Big/ \frac{n(n-1)}{2}.$$

The density of a network lies between 0 and 1, with 0 being the empty network of no edges and 1 if there is an edge between all pairs of nodes.

Representing by $A_{ij} \geq 0$ an edge between nodes i and j, we can describe the entire network using its adjacency matrix $\mathbf{A} = (A_{ij})_{i,j=1,\ldots,n}$. We call a graph *binary* if two nodes i and j are either connected ($A_{ij} = 1$), or not ($A_{ij} = 0$). Figure 1.1a illustrates how to convert a binary network into an adjacency matrix for a toy example. We will see in Chapters 3 and 4 that adjacency matrices make generalizations of graphs easy and are useful for the analyses of networks.

Each of the networks in the introduction can formally be described as a graph. We group these networks by the nature of their relationships. A prominent binary graph is a *simple* network where we assume in addition to $A_{ij} \in \{0, 1\}$: the relationships are symmetric ($A_{ij} = A_{ji}, \forall i, j$); and there are no self-loops, i.e., a node cannot connect to itself ($A_{ii} = 0, \forall i$). Friendship networks for instance are often modeled as simple graphs. Networks where two nodes can have more than one edge are called *multi-edge* networks; e.g., an email interaction network. When the connections between nodes i and j are quantified with a weight we call the network *weighted*; and networks where the relationships are not symmetric are called *di-*

rected networks. For the scope of this thesis, we concentrate on undirected networks without self-loops (i.e., $\boldsymbol{A}$ is symmetric and $A_{ii} = 0, \forall i$), unless otherwise specified.

As illustrated in Figure 1.1b, the *degree* $d_i = \sum_{j \neq i} A_{ij}$ denotes the number of connections of node i. The degree plays a central role for this work as we will see in Chapter 3 about the degree-based model. In practice, scientists often analyze the degree sequence of an observed network, which is a vector of all degrees sorted in non-decreasing order. To discuss community structure, we partition nodes into *groups* (i.e., *communities*) as illustrated in Figure 1.1c. The function $\boldsymbol{g}$ denotes the *community assignment* of the network such that $g(i)$ denotes the group of node i.

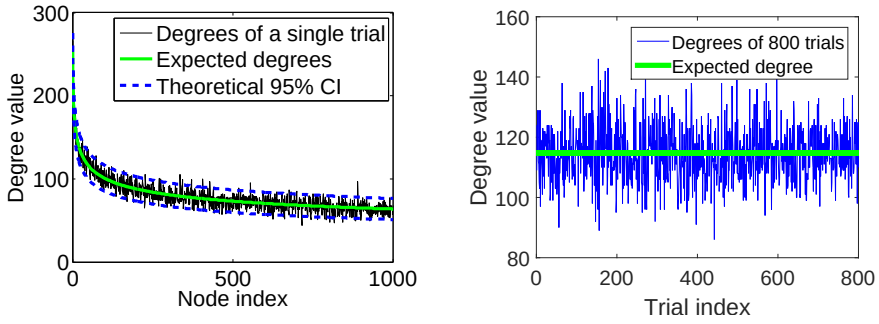
A *walk* on a graph is a sequence of alternating nodes and edges $(\nu_0, e_1, \nu_1, e_2, \nu_2, \ldots, \nu_l)$; where the edge e_{i+1} between nodes ν_i and ν_{i+1} needs to be present in the network for $i = 0, \ldots, l$. The *length* of this walk is said to be l. A *cycle* is a walk of length at least three that starts and ends at the same node but does not pass through any other node twice. A *path* is a walk without repeated nodes and edges. The *distance* between two nodes is the length of the shortest path connecting them where for weighted networks we calculate the sum of the weights. The *diameter* of a graph is the longest distance between any two nodes in the graph.

A graph is called *connected* if there exists a walk from every node to every other node. A *component* is a maximally connected subgraph; i.e., adding any other node to this subgraph would break the connectedness. The component of a graph that includes the largest number of nodes is called the *largest component*. A graph where there is an edge between every two nodes is called *complete* and a complete subgraph is called a *clique*. In *regular* graphs, every node has the same degree.

1.2.2 Concepts

In this work, we discuss *random networks* where the edges A_{ij} are random variables and we understand an observed network as a realization of a random network. In this context, there are two different types of variation: Firstly, within a single network the behavior of the nodes varies across index; e.g., Figure 1.2a displays the degrees of all nodes of the same network. Secondly, when we draw several independent and identically distributed (iid) replicates from the same network model, the behavior of a specific node varies across trials; e.g., Figure 1.2b displays the degree of the 53^{rd} node across 800 replicates. Both types of variation will be addressed in this work.

The results of this thesis are based on the analysis of how properties of a network



(a) In a single network across nodes.

(b) For the single node 53 across trials.

Figure 1.2: Two types of variation in networks: the degrees in simulated networks with $1,000$ nodes from a power law model ($p_{ij} = 0.81(ij)^{-0.2}$ for all $i < j$).

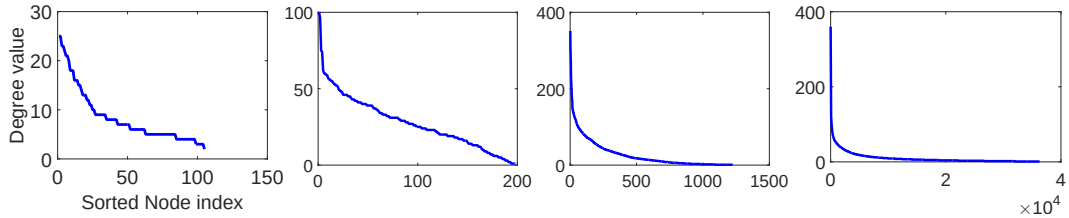
change when the number of nodes n grows. Formally, we consider a *sequence of networks* $\{G^n = (V^n, E^n)\}_{n\in\mathbb{N}}$ with $|V^n| = n$ and analyze how the properties of G^n change as $n \to \infty$. For instance, let us consider a student friendship network. Would we expect students in larger schools to have more friends? If we keep increasing the size of the school, will the number of friends keep increasing? Researchers agree that there is an upper limit to the number of people we can have an active social relationship with, the *Dunbar's number* [46, 146]. In network terms, Dunbar's number is an upper bound of the degree of every node. Note that all network properties may depend on n, e.g. the degree d_i^n of node i, but for notational convenience we omit the superscript n.

1.3 Properties of networks

To analyze networks, we first need to describe them. In this section, we introduce several network properties that have frequently been observed in practice. These properties either address the local structure where we focus on connections within relatively small neighborhoods across the network; or the global structure where we identify statements that hold for the entire network. We now discuss several local and global properties consecutively.

1.3.1 Local properties

In many networks, we observe *degree heterogeneity*: there is a large variability between the degrees of the nodes of the same network; with some nodes having in order of magnitude more connections than the average degree. This phenomenon is often coined the *scale-free*



(a) Political books (b) Jazz bands (c) Political weblogs (d) Co-author network

Figure 1.3: Degree versus sorted node index: visualization of the power law behavior of the degree sequences of four binary networks.

behavior. In particular, in many networks the expected proportion of nodes with $d_i = k$ scales approximately as $k^{-\beta}$—the *power law* behavior of the degree sequence [10]. Note that for many networks this holds only for the majority of the degrees, the exception being the degrees of lower value [32]. The power law behavior has been observed for instance for the internet, and social, and citation networks with β typically varying between 2 and 3 [48, p. 11]. Figure 1.3 displays the sorted degrees of four networks: a network of books [108] where books are connected if they have frequently been purchased together; a network of jazz bands [59] where bands are connected if they have at least one band member in common; a network of political commentary websites (weblogs) [1] where weblogs are connected if they refer to each other; and a network of physicists [104] where physicists are connected if they have co-authored a manuscript. For more details on these datasets see Chapter 5.

Node centrality is a measure for the importance of a node in a network. In a social network, the most central person will be best to spread information or crucial to prevent a disease from spreading [28]. In economics, Diebold and Yilmaz use centrality (there called connectedness) to assess the risk attached to the default of economic institutions [42]. In a gene regulatory network, the centrality of a gene indicates how lethal its deletion would be [70]. There are many different centrality measures but they often build up on the following four: degree, closeness, betweenness, and eigenvector centrality; which we now introduce.

The *degree* centrality of a node is measured by its degree

$$c_D(i) = d_i,$$

and reflects the concept that the importance of a node is well described by the number of its direct connections. Since the degree is of importance to the work here, so is the degree centrality; as we will see in Chapter 3 about the degree-based model. Furthermore, Zerubavel et al. [152]

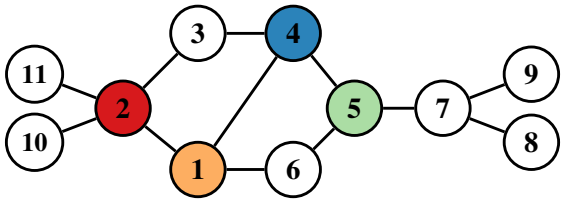


Figure 1.4: Toy example to illustrate the difference between four centrality measures [79]. The coloring indicates the most central node according to the degree (red, node 2), closeness (blue, node 4), betweenness (green, node 5), and eigenvector centrality (orange, node 1).

identify brain regions that relate to affective valuation and social cognition in humans conducting a fMRI study. The authors use the degree centrality as the base line measure for popularity of individuals. Paluck et al. [118] implement an intervention study in a social network in 56 schools to test whether promoting positive behavior in a few nodes might be sufficient to change the behavior of the majority of the network (as mentioned above in Section 1.1). The authors report that the spread of community social norms is the most effective when the intervention is applied to the "social referents"—community members with high degree.

The *closeness* centrality captures the notion that a node is central if it is closely connected to many other nodes; thereby taking into account more than the direct neighbors. The closeness centrality of node i is measured as the inverse of the distance (denoted by "dist") of node i to all other nodes [128]:

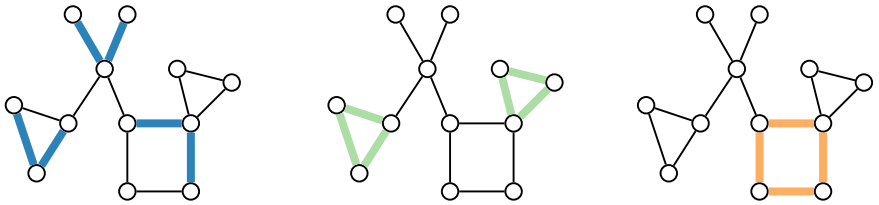
$$c_{CL}(i) = \frac{1}{\sum_{j=1}^{n} \mathrm{dist}(i,j)}.$$

Mathematically, this distance is only defined in connected graphs. To circumvent this issue, we may report the closeness centrality conditioned on the largest component or for each component individually.

The *betweenness* centrality measures how many paths go through a node. For instance if an edge represents a communication, the number of paths that go through a node counts how often a node can influence the information spreading through the network. With $s(j,l|i)$ denoting the number of shortest paths between nodes j and l passing through i, Freeman [57] defines betweenness centrality of node i as

$$c_B(i) = \sum_{j \neq l \neq i} \frac{s(j,l|i)}{\sum_{m=1}^{n} s(j,l|m)}.$$

If the shortest path is unique for all pairs of nodes, $c_B(i)$ counts the number of shortest paths that pass through i.



(a) Three 2-stars (b) All triangles (c) All 4-cycles

Figure 1.5: Toy example to illustrate network motifs. We display 2-stars: subgraphs with three nodes and two edges; triangles: complete subgraphs with three nodes; and 4-cycles: cycles of length four. Note, in Figure 1.5a we only highlight three out of many 2-stars for better visibility.

Eigenvector centrality captures the notion of "prestige" where a node's importance is judged by the importance of its neighbors. It is typically measured as a function of an eigenvector of a linear system of equations related to the adjacency matrix. One of many examples of eigenvector centralities is defined by Bonacich [23] as

$$c_{eig}(i) = \frac{1}{\alpha} \sum_{(i,j)\in E} v(j),$$

where α denotes an eigenvalue of the adjacency matrix $\boldsymbol{A}$ and $\boldsymbol{v}$ the corresponding eigenvector. Bonacich recommends to choose α as the largest eigenvalue.

Figure 1.4 illustrates that although all four measures address centrality, they in fact measure different quantities. These centrality measures may have different interpretations in practice. For instance in protein interaction networks, pharmacologists are interested in identifying proteins that are correlated with gene expression dynamics, but that are not lethal; such that a medication can target these specific proteins. Both betweenness centrality and degree centrality are associated with gene expression (betweenness centrality stronger than degree centrality) and lethality. However, nodes of medium to low degree centrality with a high betweenness centrality tend to be less likely to be lethal than the average protein [70].

In many networks, we observe that nodes tend to gather in "small neighborhoods", such that they have more connections within their neighborhood than on average to all other nodes. One way to quantify this is the *clustering coefficient* of a graph G that measures the proportion of 2-stars that close to form triangles. As illustrated in Figures 1.5a and 1.5b, 2-stars are subgraphs with three nodes and two edges ($\angle$); and triangles are complete subgraphs with three nodes ($\triangle$). To be more precise, the clustering coefficient $cl \in (0, 1)$ is defined as

$$cl(G) = \frac{1}{|V'|} \sum_{i\in V'} \frac{\text{count}_{\triangle}(i)}{\text{count}_{\angle}(i)},$$

where $\mathrm{count}_{(\cdot)}(i)$ denotes the count of occurrences of $\cdot$ centered at node i; and V' denotes the set of all nodes with at least two connections. In a network with a high clustering coefficient there is a strong tendency for 2-stars to form triangles. This phenomenon is also called *transitivity*. It is often observed in social sciences and commonly interpreted as "friends of friends tend to be friends" [112].

Comparing networks across sciences, we observe not only triangles but a variety of small subnetworks called *motifs*. The difference between two networks can be quantified by counting the occurrences of these motifs [100, 112]. For instance, for a subnetwork with three nodes there are two different motifs: 2-stars and triangles (see Figures 1.5a and 1.5b). While in social sciences we often observe transitivity, in gene regulatory networks and neuronal networks we observe in addition to transitivity an increased number of 4-cycles [100]. As illustrated in Figure 1.5c, 4-cycles are cycles of length four. We will return to the topic of motif counts in Section 2.3 in Chapter 2 on prominent challenges in network modeling.

1.3.2 Global properties

Stepping away from the node-centric view, we now discuss several network properties that are global. In many sciences, e.g. neuroscience, we observe functional units in networks that can be described as a *community structure* and may be used to reduce the dimensionality and noise of the data [36] (see Chapter 1.1). Formally, a community structure is a partition of the nodes into communities. Since we can partition the nodes arbitrarily, we denote a community structure as "informative" or "assortative" when there are more edges within communities than across communities. An informative community structure reflects a different aspect of the phenomenon described above that nodes tend to gather in small neighborhoods. In social sciences, this phenomenon is termed "homophily" and implies that connections between people of the same community occur at a higher rate than between communities because of the similarity of people of the same community [98].

The Harvard sociologist Stanley Milgram coined the term *small-world property* in 1967 when he conducted a study suggesting that every two people on this planet are only separated by at most six other people [99]. The fascination of this example results from the fact that this distance between two nodes is much smaller than we would expect at random (i.e., if all edges were placed uniformly at random). The small-world property was therefore formalized by Watts and Strogatz as a small average distance and a high clustering coefficient. While Milgram only studied social networks, Watts and Strogatz show that the small-world property in fact holds

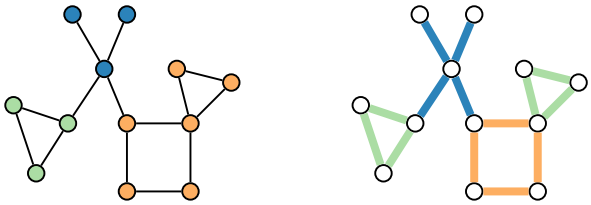
for networks in many scientific fields; e.g., neural networks in neuroscience, power grids in electrical engineering, and collaboration networks in social sciences [145]. Formally, we talk about a small average distance when in a sequence of networks the expected average distance between two nodes scales as $\mathcal{O}(\log n)$ [145]. The order notation is explained in Appendix A.1. In practice, scientists address this asymptotic property by computing the average distance of the largest component in the observed network (often a single snapshot) [79]. Watts and Strogatz introduce a generating model that leads to networks possessing the small-world property which we will discuss in more detail in Section 1.4.2.

Many properties of network models depend on the *sparsity* of the network: the relation between the number of edges and the number of nodes [20, 87]. For instance, let us assume a random network where an edge between every two nodes occurs equally likely with probability c/n (see Erdős-Rényi graphs in Section 1.4.1). The constant c strongly influences the number of edges in relation to the number of nodes and in other words, the sparsity of the network. At the same time, the value of c determines whether the network is connected: Consider a sequence of networks where $0 < c < 1$. Then, the largest component includes with high probability at most $\mathcal{O}(\log n)$ nodes, leading to a network that is not connected. In contrast, if $1 < c$ the largest component includes with high probability almost all nodes $(\Theta(n))$. For an explanation of the notation see Appendix A.1. This phenomenon was first described by Erdős and Rényi and is commonly referred to as the "emergence of the giant component". For an overview about the work on the giant component see [20].

The definition of sparsity varies depending on the context. For this thesis, we use the definition by Bollobás and Riordan [21, 22]. Recall that n denotes the number of nodes and N the number of edges. Then, a sequence of networks is

$$\begin{aligned} \text{dense: } & N = \Theta\big(n^2\big), \\ \text{sparse: } & N = o\big(n^2\big) \text{ but } N = \omega(n), \\ \text{extremely sparse: } & N = \Theta(n). \end{aligned}$$

For definitions of Θ, o, and ω see Appendix A.1. In practice, we say a network is dense, sparse or extremely sparse if it seems reasonable to assume the observed network is a snapshot of such a sequence of networks. Bollobás and Riordan call networks with $o(n)$ edges (more sparse than extreme sparse networks) to be below the minimum sensible density since the average degree of a node would asymptotically go to 0. Most observed networks are sparse or extremely sparse [90].



(a) Node dependence (b) Edge dependence

Figure 1.6: Toy example to illustrate the difference between node-dependent and edge-dependent network models. The coloring is defined by covariate in Figure 1.6a, and by motif in Figure1.6b. In both figures, things of the same color are modeled as stochastically equivalent.

1.4 Modeling of networks

So far, we have focused on the descriptive analysis of networks. To proceed to inference and prediction, we first need to introduce statistical models for networks. A statistical model $\mathcal{M}$ for networks is a set of probability distributions $\Pr_{\boldsymbol{\theta}}$ on the adjacency matrix $\boldsymbol{A}$; i.e., with Ψ defining the set of all parameters $\boldsymbol{\theta}$, and Λ the set of all adjacency matrices, we obtain

$$\mathcal{M} = \{\Pr_{\boldsymbol{\theta}}(\boldsymbol{A}) : \boldsymbol{A} \in \Lambda, \boldsymbol{\theta} \in \boldsymbol{\Psi}\}.$$

The nature of the relationships determines the set Λ. For instance, in a multi-edge network the entries of the adjacency matrix $\boldsymbol{A}$ are counts, so we obtain $\Lambda = \mathbb{N}_{\geq 0}^{n \times n}$. In a weighted network, in contrast, it follows that $\Lambda = (0, 1)^{n \times n}$. We often have additional assumptions, e.g. in a simple network the edges are undirected and there are no self-loops: $\Lambda = \left\{\boldsymbol{A} \in \{0, 1\}^{n \times n}; \boldsymbol{A}^T = \boldsymbol{A}, A_{ii} = 0 \; \forall i\right\}$.

As always, by modeling networks we encounter a trade-off between fit and complexity. If we model the network with as many parameters as there are possible edges (i.e., $n(n-1)/2$), we achieve perfect fit but have not gained any insights (i.e., overfitting). Instead, we reduce the dimensionality where there are two fundamentally different approaches illustrated in Figure 1.6. First, one assumes node dependence: all edges are independent given the nodal attributes (see Figure 1.6a). For instance in email communications in companies, employees of the same department tend to communicate more with each other than with other departments (see data analysis in Section 5.3). Second, one assumes edge dependence: the probability distribution $\Pr_{\boldsymbol{\theta}}$ only depends on the relation of the edges, independent of which nodes are involved (see Figure 1.6b). For instance in social networks, we often observe an increased number of triangles since friends of friends tend to be friends (as mentioned in Section 1.3.1 on local properties).

We now discuss models of both approaches.

1.4.1 Node dependence

Since our work is focused on community structure, we categorize the models here into two types: those with a lack of community structure and those that support community structure.

Models with a lack of community structure

The *Erdős-Rényi* graph $G(n, p)$ is the simplest and probably most studied model for simple networks [49, 58]. It assigns an edge to each pair of distinct nodes independently with the same probability $p \in (0, 1)$. It thereby models all nodes as stochastically equivalent; e.g., all nodes have the same expected degree. Thus across nodes, there is a lack of degree heterogeneity compared to what is frequently observed in practice (see Section 1.3).

To incorporates degree heterogeneity, the *degree-based* model allows for diverse propensities to connect across nodes [31]. It assigns each node a single nonnegative weight w_i to model the success probability of an edge between nodes i and j in a binary network as

$$p_{ij} = \frac{w_i w_j}{\|\boldsymbol{w}\|_1}, \text{ with } \|\boldsymbol{w}\|_1 = \sum_{i=1}^{n} w_i.$$

All edges are assumed to be conditionally independent given the parameters. To ensure that $0 \leq p_{ij} \leq 1$, the weights are constrained to $\max_l w_l^2 < \|\boldsymbol{w}\|_1$. If self-loops are allowed, the expected degree of node i is equal to its weight w_i and the model is therefore often referred to as random graph with given (expected) degrees.

A special case of this model are *power law networks* where $w_i \propto i^{-\gamma}$ with $\gamma \in (0, 1)$, introduced to match the power law behavior of degree sequences often observed in practice (see Section 1.3). Chung and Lu [31] compute the average distance and diameter of networks generated by the degree-based model, and show that for the special case of power law networks we obtain the small-world property. To be more precise, for a sequence of networks with $\gamma \in (0.5, 1)$—a range often observed in practice [10, 31, 32, 48]—the average distance is almost surely $\mathcal{O}(\log \log n)$ [31].

To fit the degree-based model, Perry and Wolfe [119] estimate the edge probabilities as

$$\hat{p}_{ij} = \frac{d_i d_j}{\|\boldsymbol{d}\|_1}.$$

Recall that d_i denotes the degree of node i. The authors show that this estimator is a near-maximum likelihood estimator in the sparse graph regime ($p_{ij} = o(1)$). The authors provide

explicit constants to check the sparsity assumption and give upper bounds for the relative error of the estimator in the least sparse case. The upper bound improves as the network becomes more sparse.

Olhede and Wolfe [116] analyze the degree distribution under the degree-based model; both for power law networks, and for iid random weights. The authors change the parameterization such that the constraints may be specified for each node separately to

$$p_{ij} = \pi_i \pi_j \text{ with } \pi_i = w_i / \sqrt{\|\boldsymbol{w}\|_1} \in (0, 1);$$

and thus, the edge probabilities are decoupled. In particular, for the special case of power law networks, the authors derive a central limit theorem for the estimators $\hat{\pi}_i = d_i / \sqrt{\|\boldsymbol{d}\|_1}$.

The *configuration model*, discussed extensively in the physics and mathematics literature, is strongly related to the degree-based model: It fixes the actual degrees to then generate the network at random with respect to the given degrees [114]. In contrast, the degree-based model fixes the expected degrees instead.

Models supporting community structure

One of the simplest models for community structure is the *stochastic blockmodel* [68]: a generalization of the Erdős-Rényi graph. The nodes are partitioned into K subsets, called blocks, and the probability of an edge between nodes i and j only depends on the group membership $\boldsymbol{g} \in \{1, \dots, K\}^n$ (see Section 1.2.1):

$$p_{ij} = \omega_{g(i),g(j)}.$$

It thereby models all nodes of the same community as stochastically equivalent. The authors assume the group membership to be known a priori and derive a straightforward maximum likelihood estimator for the edge probabilities as the sample proportion. In the remainder of the article the authors introduce a generalization of the stochastic blockmodel to directed networks. Aicher et al., Airoldi et al., and Latouche et al. provide generalizations of the stochastic blockmodel to weighted networks and mixed and overlapping memberships, respectively [2, 3, 82]. We will come back to the stochastic blockmodel under the assumption of unknown (latent) communities in Section 2.4 on community detection.

As the Erdős-Rényi graph, the stochastic blockmodel lacks heterogeneity of degrees. The *degree-corrected stochastic blockmodel* [75] marries the concepts of community structure and degree heterogeneity by combining the stochastic blockmodel and the degree-based model. The

authors assume a multi-edge network and model the expected edge counts $\mathbb{E}\, A_{ij}$:

$$\mathbb{E}\, A_{ij} = \pi_i \pi_j \omega_{g(i),g(j)}$$

where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ are the node-specific parameters and $\omega_{g(i),g(j)}$ only depends on the group membership $\boldsymbol{g}$. In contrast, all models presented so far assume binary networks and thus model edge probabilities p_{ij} instead. The authors suggest a two-step procedure for model fitting: a profile likelihood estimation for the parameters $\boldsymbol{\pi}$ and $\boldsymbol{\omega}$ conditioned on $\boldsymbol{g}$, and then a heuristic algorithm to identify the optimal group membership $\boldsymbol{g}$. The authors avoid maximizing the likelihood over all possible group memberships since that is a NP-hard problem [24].

The *latent space models* introduce a latent (unobserved) space where each individual node i has an unobserved ("latent") position $\boldsymbol{z}_i \in \mathbb{R}^d$ [67]. The model assumes that nodes that are close in the latent space and have common covariates are more likely to connect. The latent space is commonly chosen to be of lower dimension; i.e., $d < n$; adding parsimony and interpretability. Additional benefits are that the latent space model incorporates both local and global structure, and transitivity, and that it outputs a meaningful visualization [131]. The authors model the edges A_{ij} as conditionally independent given the latent positions $\boldsymbol{z}_i$ and the covariates $\boldsymbol{x}_{ij}$ using a logistic regression model; i.e., with $\boldsymbol{\beta}$ denoting all parameters

$$\operatorname{logit} p_{ij} = \beta_0 + \boldsymbol{\beta}\boldsymbol{x}_{ij} - |\boldsymbol{z}_i - \boldsymbol{z}_j|.$$

The latent positions $\boldsymbol{z}_i$ are modeled using diffuse independent normal priors: $z_{ij} \overset{iid}{\sim} \mathrm{Normal}(0, 100)$. The corresponding log-likelihood as a function of the latent positions is not concave; and much caution must be taken to differentiate local from global maxima. The authors suggest Markov chain Monte Carlo algorithms to infer the latent positions and derive confidence regions. The latent positions provide a soft clustering of the nodes into d clusters while regressing on covariates. The latent space model was extended to include random node-specific effects [66] using a mixed effects model (a generalized linear model with structural error terms); and community structure [63] by modeling the latent positions as a mixture of Normal random vectors (see Section 2.4 on clustering in networks). Krivitsky et al. [81] combine all four effects: homophily based on common covariates, transitivity, random node-specific effects, and community structure into a single model. To test for the dependence between latent structure and covariates, and to predict missing values, Fosdick and Hoff [53] model the covariates and latent structures as random simultaneously.

Closely related to the latent space model is the *random dot product graph* [66, 151]. As for the latent space model, each node has a latent position $\boldsymbol{z}_i \in \mathbb{R}^d$, where all z_{ij} are modeled as *iid*

random variables. In contrast to the latent space model, it is assumed that $\boldsymbol{z}_i^T \boldsymbol{z}_j \in (0,1)$ for all i, j and the probability for an edge A_{ij} is modeled as the inner product of the latent positions:

$$p_{ij} = \boldsymbol{z}_i^T \boldsymbol{z}_j.$$

The random dot product graph exhibits scale-free behavior of the degree sequence and the small-world property (small average distance almost surely and a high clustering coefficient) [151]. Sussman et al. [138] introduce an estimator for the latent positions $\widehat{\boldsymbol{Z}} = (\hat{\boldsymbol{z}}_1, \ldots, \hat{\boldsymbol{z}}_n)^T \in \mathbb{R}^{n \times d}$. Let $\boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$ be the eigen-decomposition of $\left(\boldsymbol{A}\boldsymbol{A}^T\right)^{1/2}$, and define $\boldsymbol{S}_{[d]} \in \mathbb{R}^{d \times d}$ the diagonal matrix of the d largest eigenvalues and $\boldsymbol{U}_{[d]} \in \mathbb{R}^{n \times d}$ the matrix of the corresponding eigenvectors. Then, assuming we know the dimension d of the latent space we may estimate the latent positions as

$$\widehat{\boldsymbol{Z}} = \boldsymbol{U}_{[d]} \boldsymbol{S}_{[d]}^{1/2}.$$

The authors show weak consistency of the estimators: $\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i\|_2^2 = o_P(1)$ for each i with $\|.\|_2$ being the Euclidean norm. Furthermore, the scaled residuals between estimated and true latent positions converge in distribution to a mixture of multivariate Normal distributed random vectors [9]. For more results on the eigen-decomposition of matrices related to $\boldsymbol{A}$ see Section 2.4.2.

Under the assumption that all parameters are random, all network models stated above can be joined to a single class of models [15]: A *graphon* defines a limiting object for simple random networks when the number of nodes goes to infinity. Let us assume that all nodes are exchangeable: the edge probabilities are invariant to relabeling of the nodes. Since then the adjacency matrix (in the limit) $\{A_{ij}\}_{i,j=1}^{\infty}$ is an exchangeable infinite array of random variables, we know that it admits a representation in functions $f(\xi_i, \xi_j, \alpha)$ with $\xi_i, \xi_j, \alpha \overset{iid}{\sim} \text{Uniform}(0,1)$ (Aldous-Hoover theorem [4, 69]) and that this representation is unique up to measure-preserving transformations [41]. To allow for sparse networks, it is common practice to multiply the graphon by a scaling factor $\rho_n > 0$ that depends on n. Thus, we obtain the following model for exchangeable simple random networks:

$$\begin{aligned} A_{ij} | p_{ij} &\sim \text{Bernoulli}(p_{ij}), \\ p_{ij} = \rho_n f(\xi_i, \xi_j, \alpha), \quad \xi_i, \alpha &\overset{iid}{\sim} \text{Uniform}(0,1). \end{aligned} \tag{1.1}$$

Since $\mathbb{E}\, A_{ij} = \int\int p_{ij} d\xi_i d\xi_j = \rho_n$, the parameter ρ_n reflects the overall sparsity of the network. Each observed $(n \times n)$ network is then a sub-network of the infinite-dimensional network modeled in Eq. (1.1). This class of models is closely related to exchangeable random graph models [4, 69, 90] and inhomogeneous random graphs [20].

The function f does not uniquely determine the probability density function of the network [14]. However, it is common to interpret a graphon instead as an equivalence class that includes f and all its measure-preserving transformations. Olhede and Wolfe [117] present a method to fit a graphon: the network histogram approximates the potentially smooth graphon by a piecewise constant generating function, the stochastic blockmodel, in the same way as a histogram approximates a probability distribution function or the Riemann sum approximates the integral of a continuous function.

Li et al. [89] incorporate network structure into *classic regression* by introducing a penalty to encourage network cohesion. Assume we are interested in the influence of p covariates $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in \mathbb{R}^n$ on an outcome $\boldsymbol{y} \in \mathbb{R}^n$, and we know the binary network connections of the participants (i.e. the adjacency matrix $\boldsymbol{A} \in \{0,1\}^{n \times n}$). Denote $\boldsymbol{\alpha} \in \mathbb{R}^n$ a node-specific effect, ϵ an error term and let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$. Then the authors suggest to model $\boldsymbol{y}$ as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\alpha} + \epsilon.$$

To estimate the coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$ of the association of the covariates with the outcome, the authors minimize a penalized residual sum of squares: with λ denoting a tuning parameter,

$$L(\alpha, \beta) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\alpha}\|_2^2 + \lambda \sum_{\forall i,j : A_{ij}=1} (\alpha_i - \alpha_j)^2.$$

Under some regularity conditions, the authors provide upper bounds for the mean squared errors for both corresponding estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$; and show that $\hat{\boldsymbol{\beta}}$ is consistent.

1.4.2 Edge dependence

We here provide two examples of models that exploit edge dependencies; one with a lack of community structure and one that supports community structure. The aim is to give a brief introduction to an alternative modeling approach; different from all models used in this work.

A model with a lack of community structure

The *Watts-Strogatz* model [145] is designed to possess the small-world property: a small average distance and a high clustering coefficient. Watts and Strogatz begin with a network where each node is connected to its r nearest neighbors (i.e., a r-regular graph); a network that tends to have very high transitivity. They then randomly reconnect each edge with probability p (while avoiding self-loops and multi-edges); introducing "short-cuts" that reduce the average distance in the graph. As a consequence, the model is a hybrid between a regular and an Erdős-Rényi

graph, for which even for small p we observe the small-world property. The Watts-Strogatz model is of particular importance to the analysis of information propagation and epidemic spread, since it allows for information to be transmitted quickly through the entire network only based on neighbor-to-neighbor communications. A discussion about epidemic spread and related work is beyond the scope of this thesis. In contrast to our work, these approaches assume the network to be non-random and analyze a random process on the network.

A model supporting community structure

The *exponential random graph* model (ERGM) [54, 144] builds up on the concept of exponential families and thereby naturally extends statistical regression to random networks. It models a simple network $G = (V, E)$ in terms of a set $\mathcal{H}$ of motifs:

$$f(\boldsymbol{A}|\boldsymbol{\theta}) = \frac{1}{\kappa} \exp\left(\sum_{H \in \mathcal{H}} \theta_H c_H \right), \tag{1.2}$$

with c_H being a count of how often the motif H occurs in the network $\boldsymbol{A}$; and the standardization constant $\kappa = \sum_{\boldsymbol{A} \in \Lambda} \exp(\sum_H \theta_H c_H)$. Originally, ERGMs included star and triangle counts (see Figure 1.5). The formula in Eq. (1.2) implies that the density factorizes over the subgraphs $H \in \mathcal{H}$. For instance if $\mathcal{H}$ includes only the edge subgraph, we assume all edges to be *iid*. For more details on the independence assumptions on f implied by $\mathcal{H}$ see the Hammersley-Clifford theorem [12].

The main advantage of the ERGM is that it can represent a variety of structural tendencies, such as transitivity, and that its parameters are easy to interpret. However, fitting the model has proved to be challenging because of three main reasons. First, computing the standardization constant κ is computationally infeasible for medium to large networks ($n > 30$) [62] since κ is a sum over the entire sample space Λ. Second, very different values of $\boldsymbol{\theta}$ can lead to essentially the same distribution f [27]. Third, ERGM models are degenerate: they put the majority of mass on the empty graph, the complete graph or a mixture of the two [62]. To overcome the degeneracy, Handcock [62], Hunter and Handcock [72], and Snijders et al. [136] introduce priors that restrict the parameter space to graphs that are neither empty nor complete. Chatterjee and Diaconis [27] deliver limiting results that identify when degeneracy occurs; and show that those graphs that are not degenerate often are indistinguishable from an Erdős-Rényi graph in the limit (i.e. $f(\boldsymbol{A}|\boldsymbol{\theta}) \xrightarrow{P} f(G(n,p))$ as $n \to \infty$). Since computing κ is time consuming even for moderately small networks, Chatterjee and Diaconis [27] provide analytical formulas for κ based on limit theorems; with the major limitation that it holds only for dense networks. Due to

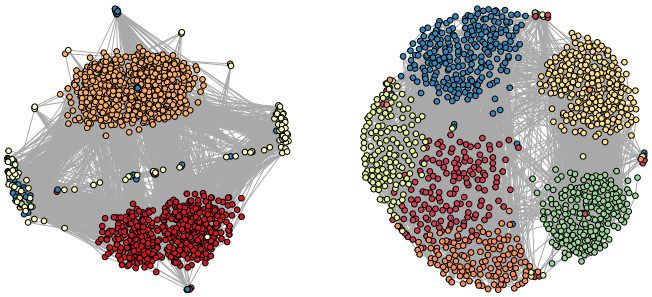
these limitations, many researchers question the suitability of ERGMs for statistical inference on networks [27].

1.5 Contributions of the thesis and their context

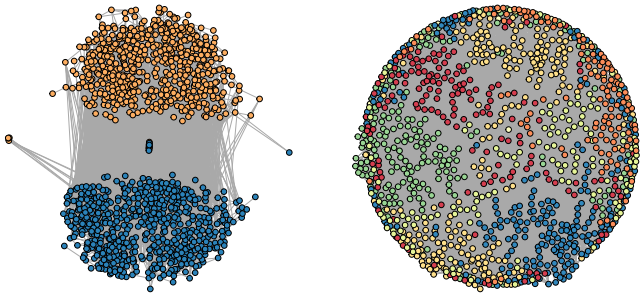
After establishing a framework in the previous sections, we now can explain the original contributions of the thesis, and embed them in the wider context of networks. We establish a methodology to identify the key characteristics that determine a network's structure. To do so, we firstly characterize a family of flexible, nonparametric models that naturally generalize the degree-based model mentioned in Section 1.4.1. Under such models, we secondly derive the theoretical foundation for modularity: an intuitive and practically effective measure of the strength of community structure; enabling us to decide which of the characteristics reflect the structure of the interactions in a network.

First, we generalize the degree-based model to a broad class of models including weighted, multi-edge, and power law networks. We fit such a model using the canonical estimator $\widehat{\mathbb{E}\, A_{ij}} = d_i d_j / \|\boldsymbol{d}\|_1$ for which Perry and Wolfe [119] show that it is close to the maximum likelihood estimator for the special case of simple networks. We show that $\widehat{\mathbb{E}\, A_{ij}}$ is weakly consistent for $\mathbb{E}\, A_{ij}$ and derive its asymptotic distribution under this broad class of network models. Our results generalize work by Olhede and Wolfe [116] who show the asymptotic distribution for the special case of power law networks. All approaches and results presented above assume the edges to be either Bernoulli or Poisson distributed. In contrast, we take a nonparametric approach: using a single parameter per node, we model only the expectation of each edge. This allows for individual node-specific differences but avoids specific distributional assumptions on the edges. Our results therefore apply to a broader class of network models, allowing us to treat (among others) power-law networks, weighted networks, and those with multiple edges.

Second, we extend the concept of community structure to reflect the complexity of observed networks. Scientists inevitably observe not only network nodes and their connections, but also additional information in the form of covariates. Many analysis approaches fail to exploit this information when attempting to explain network structure, and instead solely focus on identifying a single "best" community structure (e.g. latent space models [67, 53], stochastic block models [68] and degree-corrected stochastic block models [75]; see Section 1.4.1). In recent works, researchers have started to use covariates to improve community detection (degree-corrected stochastic blockmodels [110], latent space models [67], modularity [153],



(a) Race (b) Year in school



(c) Gender (d) Randomized

Figure 1.7: A student friendship network illustrated for four different community assignments, each defined by a covariate [53, 110, 124]; implemented using igraph [38].

ERGMs [50], see Section 2.4.1). However, the concept of a single best community assignment leads to a loss of interpretability and presents a barrier to understanding. We solve this problem, by acknowledging that each of several community assignments may describe different aspects of a network's structure. To add interpretability, we define these community assignments based on covariates, and show how to decide which of these covariate-based community assignments leads to a valid summary of the network structure. In the student friendship network shown in Figure 1.7, for example, this means we can evaluate whether communities based on common gender, race, or year in school can explain the observed structure of the friendships.

Technically, we derive modularity from first principles, and give it a formal statistical interpretation. Effectively, modularity summarizes the difference between observed and expected within-community edges under a model for no community structure. We derive the large-sample distribution of modularity under the nonparametric class of degree-based models, enabling us to compute a p-value for the significance of a covariate-based community structure. This provides for the first time an objective measure of whether or not a particular value of modularity is meaningful. As a result, we deliver a flexible, nonparametric approach to identify those covariates that reflect a network's structure.

Papers and preprints

[56] Franke B, Wolfe PJ (2016) Network modularity in the presence of covariates. arXiv:1603.01214.

1.6 Outline

The remaining thesis is organized as follows. We first review the literature about prominent challenges in network modeling, including community detection using modularity (Chapter 2). We then present our original contributions. We extend the degree-based model to a nonparametric family and derive its asymptotic properties (Chapter 3). We establish a theoretical framework based on modularity to assess whether a covariate-based community assignment is informative for the intersections in a network (Chapter 4). Here, we show that the model underlying modularity is a degree-based model and derive a bias-variance decomposition for modularity. This enables us to establish the asymptotic distribution of modularity and to deliver a p-value reflecting the explanatory power of covariates on network connections. In Chapter 5, we turn the theory into a methodology to identify covariates that are informative for the connections in a network. After validating our method on four benchmark examples, we analyze email interactions in a multi-edge corporate email network identifying those covariates that reflect the network's structure. We conclude this thesis with a discussion of the prominent challenges; and point out possible directions of follow-up research (Chapter 6).

Chapter 2

Literature review of prominent challenges in network modeling

In this chapter, we present a literature review about the current challenges in network modeling. In Chapter 1, we have seen how to describe networks and how networks are modeled so far. Due to technological advances, the amount of data that we can store and process increases on a daily basis; allowing us to collect data that better reflect the complexity of nature. However, to actually gain insights from the data our network models must catch up to reflect the high dimensionality of the data (see Section 2.1). All network models enable us to draw inference from data, but for the results to be defensible we must be able to quantify the goodness-of-fit of network models (see Section 2.2). In science, it is common to collect more than one dataset and by understanding the agreements and differences, we gain insights on what are the driving forces. In Section 2.3, we describe the current state-of-the-art for networks in this regard. Having mentioned the complexity of the data above, clustering gives us a means to extract information from networks: identifying groups of nodes that have a stronger tendency to connect to each other than to other nodes. In Sections 2.4 and 2.5, we discuss different approaches for clustering in networks with a special emphasis on modularity because of its popularity and its importance for the thesis. Our review of the current challenges in network modeling is not exhaustive but we rather focus our discussion on the four topics motivated above that we find the most prominent.

2.1 Network models with higher dimensionality

We now review models that have an increased dimensionality compared to the more traditional models in Section 1.4. Modern technology enables us to collect data that are increasingly large

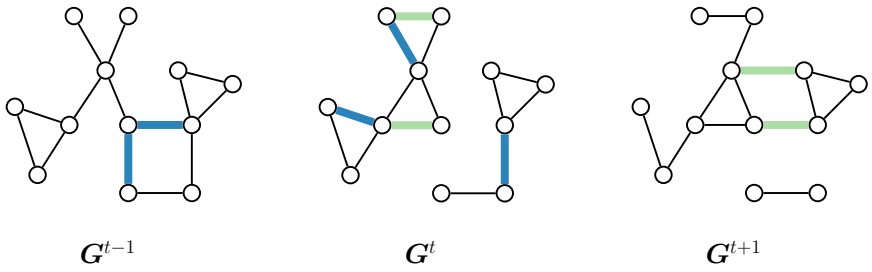


Figure 2.1: Toy example to illustrate dynamic networks. Green and blue mark edges that appear in this time step or will disappear in the next time step, respectively. The node set is fixed.

and diverse in structure—already in 2012 it was estimated that data collection was growing at 50% per year [55, 91]. Increasing the dimensionality of our models enables us to incorporate more of the complexity of the network data. The two leading approaches combine several networks into a single model. One assumes there to be a temporal dependence between the networks and the other allows for an arbitrary dependence in the spirit of a network of networks. We now discuss both approaches.

2.1.1 Dynamic networks

As illustrated in Figure 2.1, *dynamic network models* are a natural extension of the models described in Chapter 1.4 where networks are altered over time. For example, in social networks, friendships are made and lost over time, and in biological neural networks the activation of neurons is time-correlated; just to name a few. Formally, we assume a time series of networks

$$\{G^t : t = 1, \dots, T\} \quad \text{with} \quad G^t = (V, E^t), \text{ and } \{\boldsymbol{A}^t : t = 1, \dots, T\}$$

being the series of adjacency matrices. All approaches mentioned below assume the node set V to be fixed over time; with the exception of the preferential attachment model.

One of the first statistical models for dynamic networks was introduced by Snijders in 1996 [134]. The authors model the adjacency matrix $\boldsymbol{A}^t$ and the attributes $\boldsymbol{C}^t$ of the nodes as Markov processes in continuous time $t \in [0, \infty)$. Denote t_0 the present and $t^* \in [0, t_0)$ all previous time points. Then, the authors define

$$\Pr\big(\{\boldsymbol{A}^t, \boldsymbol{C}^t : t > t_0\} | t^* \in [0, t_0)\big) = \Pr\big(\{\boldsymbol{A}^t, \boldsymbol{C}^t : t > t_0\} | t_0\big).$$

The authors assume that a change in $\boldsymbol{A}_i^t$ or $\boldsymbol{C}_i^t$ of connections and attributes of node i occurs at a rate $\lambda_i(\boldsymbol{A}^t, \boldsymbol{C}^t)$ and model the waiting time until the next change by node i as a negative exponential distribution, with expected value $1/\lambda_i(\boldsymbol{A}^t, \boldsymbol{C}^t)$. The authors further assume that each

node has an incentive to optimize its attributes and connections to its own benefit. They model the decision making of node i using a tension function $p_i(\boldsymbol{A}^t, \boldsymbol{C}^t)$ that includes both a deterministic and random component. The authors infer the parameters of the tension function and the rate of change using a method of moments based on parametric bootstrap and approximate the covariance matrix of the vector of estimators using the delta method. The authors point out that the approach is computationally expensive, lacks in statistical efficiency and caution the reader that the variance estimators might be instable.

The *preferential attachment model* describes a process of adding nodes, and connecting these to nodes with a probability proportional to their degree; leading to a "richer getting richer" phenomena and a power law degree sequence [10]. Although the preferential attachment model is understood as a single snapshot of a network, in contrast to multiple snapshots over time, it does describe a dynamic process of generating a network. In a similar way, we can describe the Watts-Strogatz model as a dynamic model (see Section 1.4.2). In contrast to the other models in this section, the preferential attachment model and the Watts-Strogatz model incorporate a change in the set of nodes but are not intended for statistical model fitting.

A generalization of the *random dot product graph for dynamic networks* is provided by Lee and Priebe [84] where the authors aim to detect change points in the behavior of weighted networks. The authors model a time series of networks with categorical edge weights in discrete-time $t \in \mathbb{N}$ as a series of random dot product graphs derived from a finite-state Markov process in continuous-time u: $\boldsymbol{W} = \{\boldsymbol{w(u)} \in \{1, \dots, d+1\}^n : u \in [0, \infty)\}$. To be more precise, for nodes i, j and edge weights l it holds for the latent positions $\boldsymbol{Z^t} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)^T \in \mathbb{R}^{n \times d}$ that

$$
\begin{aligned}
& z_{il}^t \overset{a.s.}{=} \int_{t-1}^{t} \delta_{w_i(u)=l} \, du, \quad \text{for } l = 1, \dots, d, \\
& \Pr\big(A_{ij}^t = l | \boldsymbol{z}_i^t, \boldsymbol{z}_j^t\big) = \begin{cases} z_{il}^t z_{jl}^t, & \text{for } l \neq 0; \\ 1 - \sum_{l=1}^{d} z_{il}^t z_{jl}^t, & \text{for } l = 0; \end{cases} \qquad \text{independent for all } i < j \\
& \Pr\big(\boldsymbol{A}^t = \boldsymbol{a} | \boldsymbol{w(u)}, u \leq t\big) = \Pr\big(\boldsymbol{A}^t = \boldsymbol{a} | \boldsymbol{Z}^t\big).
\end{aligned}
$$

The authors assume that the probabilities of the stochastic process $\boldsymbol{W}$ to take values in $\{1, \dots, d+1\}$ change for a small community of nodes and they aim to detect the corresponding time point. The authors introduce two approximations to make the problem analytically tractable and show that their total variation distance under a dynamic random dot product graph is asymptotically small. Durante and Dunson [47] work on a strongly related model of a dynamic random dot product graph, where the latent positions evolve in a continuous

Markov process. Due to using a logistic link between the probability of an edge and the latent positions—the main difference—the authors obtain a computationally tractable formulation. They introduce an algorithm to both infer the posterior distribution and estimate the dimension of the latent space simultaneously.

In [131], Sewell and Chen generalize the *latent space model for time-varying, simple networks* and thus, this work is strongly related to the dynamic random dot product graph by Lee and Priebe [84]. To be more precise, the authors model the latent positions $\boldsymbol{Z}^t = \big(\boldsymbol{z}_1^t, \dots, \boldsymbol{z}_n^t\big)^T \in \mathbb{R}^{n\times d}$ as a Markov process in discrete time $t \in \mathbb{N}$. Denote $\boldsymbol{I}_d$ the $(d \times d)$-identity matrix, and $\boldsymbol{\theta}$ all parameters. Then, the authors define the initial distribution of the latent positions at time $t = 1$ as

$$\pi\big(\boldsymbol{Z}^1|\boldsymbol{\theta}\big) = \prod_{i=1}^{n} \mathrm{Normal}\big(\mathbf{0}, \tau^2\boldsymbol{I}_d\big).$$

The transition probability is defined as

$$\Pr\big(\boldsymbol{Z}^t|\boldsymbol{Z}^{t-1}, \boldsymbol{\theta}\big) = \prod_{i=1}^{n} \mathrm{Normal}\big(\boldsymbol{z}_i^{t-1}, \sigma^2\boldsymbol{I}_d\big).$$

Networks at different time points are conditionally independent given the latent positions. At each time point, the edges are modeled using a latent space model: a logistic regression model with the distance in latent space being an explanatory variable and where we assume conditional independence of the edges given the latent positions and the parameters $\boldsymbol{\theta}$ ([67], see Section 1.4.1). The authors estimate the model parameters and the latent positions using Markov chain Monte Carlo methods; where they provide approximations to speed up the algorithm. As an output, the authors deliver a temporal trajectory of each node in the latent space. Furthermore, the authors address the problem of missing data, and prediction; and demonstrate their method on simulated and observed data.

Westveld and Hoff introduce a *dynamic network regression* framework where the edges are modeled as conditionally independent using a generalized linear model with mixed effects [147]. Denote s_i^t, r_i^t the node-specific effects of node i as a sender and receiver, respectively, and e_{ij}^t the residual error terms. The random effects s_i^t, r_i^t, e_{ij}^t are modeled using discrete-time Markov processes. Furthermore, denote $\boldsymbol{x}_{ij}^t$ the fixed effects, e.g., covariates, and h the link function. The authors then model the adjacency matrix at time t as

$$\begin{aligned}\mathbb{E}\big(A_{ij}^t|\theta_{ij}^t\big) &= h(\theta_{ij}^t),\\ \theta_{ij}^t &= \big(\boldsymbol{x}_{ij}^t\big)^T\boldsymbol{\beta}^t + s_i^t + r_j^t + e_{ij}^t.\end{aligned}$$

The authors provide Markov chain Monte Carlo algorithms for parameter estimation for Gaussian and binary networks and apply the method to data on international trade and militarized interstate disputes. This model partly builds up on the static, latent space model by Hoff [66], but it exchanges the latent positions against generic residual error terms.

To incorporate community structure into dynamic networks, Xing et al. [149] introduce a *dynamic mixed membership stochastic blockmodel* that generalizes the mixed membership stochastic blockmodel [2]. In the mixed membership stochastic blockmodel, a node may belong to multiple communities, each with a fractional membership; thereby combining the concepts of the stochastic blockmodel and the latent space model. Denote $\boldsymbol{B}^t = \left(\beta_{kl}^t\right)_{k,l=1,\dots,K}$ the probabilities to interact between nodes of communities k and l at time t. For the dynamic mixed membership stochastic blockmodel, the authors model for each node i the fractional membership $\boldsymbol{\pi}_i^t = \left(\pi_{i1}^t, \dots, \pi_{iK}^t\right)$ at time t:

$$\boldsymbol{\pi}_i^t \overset{iid}{\sim} \text{Logistic-Normal}\left(\boldsymbol{\mu}^t, \boldsymbol{\Sigma}^t\right).$$

Thus, $\boldsymbol{\Sigma}^t$ determines the correlations of the memberships between nodes. For each edge A_{ij}, nodes i and j get assigned to a single community with a probability according to their fractional memberships $\boldsymbol{\pi}_i^t$ and $\boldsymbol{\pi}_j^t$:

$$\boldsymbol{z}_l^t \sim \text{Multinomial}\left(\boldsymbol{\pi}_l^t, 1\right), \quad \text{for } l = i, j \text{ independent.}$$

An edge between nodes i and j is then modeled:

$$A_{ij}^t \overset{iid}{\sim} \text{Bernoulli}\left(\beta_{lk}^t\right), \quad \text{if } z_{il}^t = 1, z_{jk}^t = 1.$$

Furthermore, the expected value of $\boldsymbol{\pi}_i^t$ is modeled as dynamic with transition matrix $\boldsymbol{M}$:

$$\boldsymbol{\mu}^1 \sim \text{Normal}(\boldsymbol{\nu}, \Psi), \quad \boldsymbol{\mu}^t \sim \text{Normal}\left(\boldsymbol{M}\boldsymbol{\mu}^{t-1}, \Psi\right).$$

Here, Ψ determines the correlation between the expected fractional memberships across time. In addition, $\boldsymbol{B}^t$ is assumed to be time dependent. With Φ denoting the variance of the probabilities to interact between communities, and b a tuning parameter, we obtain

$$\beta_{lk}^1 \sim \text{Logistic-Normal}(\iota, \Phi), \quad \beta_{lk}^t \sim \text{Logistic-Normal}\left(b\beta_{lk}^{t-1}, \Phi\right).$$

The authors aim to infer the dynamics of the fractional memberships $\boldsymbol{\pi}_i^t$, for all i, t; as well as the dynamical community relations $\boldsymbol{B}^t$. Fitting the dynamic mixed membership stochastic blockmodel, we also estimate the correlations of the fractional memberships between nodes $\boldsymbol{\Sigma}^t$

and across time Ψ, and learn about the dynamics between communities Φ. Since the model is intractable the authors introduce a variational EM algorithm to approximate the maximum likelihood estimation. The authors demonstrate their method and algorithm on synthetic and observed data (beside others the Enron data, see Section 5.3).

In contrast to the dynamic models described so far, the *temporal ERGM* introduced by Hanneke et al. (2010) has a memory: the networks across time are dependent, even when conditioned on all parameters [64]. The temporal ERGM extends the ERGM based on a Markov process on the adjacency matrices. To be more precise, the initial network $\boldsymbol{A}$ is any ERGM and the transition probability for $t \in \mathbb{N}$ is given by

$$\Pr\big(\boldsymbol{A}^t | \boldsymbol{A}^{t-1}, \boldsymbol{\theta}\big) = \frac{1}{\kappa} \exp\left(\sum_{H \in \mathcal{H}} \theta_H c_H\big(\boldsymbol{A}^t, \boldsymbol{A}^{t-1}\big) \right),$$

where κ denotes the standardization constant, $\boldsymbol{\theta}$ the vector of all model parameters and $\mathcal{H}$ a set of motifs. The motif counts $c_H\big(\boldsymbol{A}^t, \boldsymbol{A}^{t-1}\big)$ here may include edges of the current and the previous network. For instance, to assess the stability of the process we may include the motif S where we count how many edges/non edges remained the same:

$$c_S\big(\boldsymbol{A}^t, \boldsymbol{A}^{t-1}\big) = \frac{1}{n-1} \sum_{i,j=1}^{n} \Big[A_{ij}^t A_{ij}^{t-1} + \big(1 - A_{ij}^t\big)\Big(1 - A_{ij}^{t-1}\Big) \Big].$$

To assess a dynamic version of transitivity, we may include the motif T:

$$c_T\big(\boldsymbol{A}^t, \boldsymbol{A}^{t-1}\big) = n \left[\sum_{i,j,l=1}^{n} A_{ij}^t A_{il}^{t-1} A_{jl}^{t-1} \right] \Bigg/ \left[\sum_{i,j,l=1}^{n} A_{il}^{t-1} A_{jl}^{t-1} \right].$$

An example for a motif of interest that only includes the current network is the density of the network at time t which captures the current sparsity. As for static ERGMs, due to the normalizing constant κ exact solutions of the likelihood maximization are computationally intractable. The authors suggest an approximate maximum likelihood approach based on Markov chain Monte Carlo sampling; demonstrate the convergence rate of the maximum likelihood estimators by simulation, and identify the properties under which the model is non-degenerate. Krivitsky and Handcock (2014) alter the parameterization of the temporal ERGM to enable estimation of the incidence: the rate at which new edges occur; and the duration: the time edges stay [80]. The formation and dissolution of edges are assumed to be conditionally independent given the previous network $\boldsymbol{A}^{t-1}$. As a consequence, the likelihood can be decomposed, and the conditional maximum likelihood fitting approach of Hunter and Handcock [72] can be generalized to the setting here. The authors apply their method to a friendship network of 26 students.

2.1.2 Multi-layer networks

The first notion of a *multi-layer network* dates back at least to the 1973 work by Craven and Wellman on networks of networks [37]. While sciences across disciplines as diverse as neuroscience [11] and social sciences [139] apply multi-layer networks, much of the current literature on methodology for multi-layer networks focuses on mathematical rather than statistical properties. A multi-layer network is defined as several layers of networks, where the nodes in each layer and across layers may be pairwise connected; and where each node may belong to several layers simultaneously. For instance, each layer might correspond to a conference with the network being the interactions of researchers; or each layer might refer to a different kind of connection: friendships, collaborations, relatives. The layers in a multi-layer network may also reflect hierarchy levels: the first for individuals, the second for families, the third for regions, and so forth. Compared to dynamic networks, it allows for more flexibility (implies less structure) since the layers may not have a natural ordering. De Domenico et al. introduce a mathematical solution to rank the nodes in a multi-layer network by their betweenness and eigenvector centralities [40]. Kleinberg et al. [78] identify a hidden geometric structure in multi-layer networks and Mucha et al. [102] address community structure. For a review on the work related to multi-layer networks see [77].

2.2 Quantifying goodness-of-fit of network models

While there is no widely accepted gold standard for the goodness-of-fit in networks yet, the parametric bootstrap is widely used. After introducing it, we describe the approaches that extend the classical goodness-of-fit measures to networks; covering AIC, BIC, the Bayes factor and a likelihood approach. The methods presented here illustrate why the goodness-of-fit for networks is so challenging while covering popular methods.

2.2.1 The parametric bootstrap

In [71], Hunter et al. introduce the *parametric bootstrap* for network models—a widely used method for the goodness-of-fit in networks—on the example of fitting an Erdős-Rényi graph and an ERGM to the Addhealth friendship data (see Figure 1.7). The authors simulate 100 networks from a model that has been fitted to the data, to then compare a set of statistics of the observed data to its empirical distributions computed from the bootstrap samples. The choice of statistics is crucial and must reflect structurally important aspects of the data. For the friendship

network, the authors look at the degree density (the relative frequency of nodes with degrees equal to $l = 1, \ldots, n-1$), the edgewise shared neighbors (the relative frequency of edges where the endpoints share l neighbors, $l = 1, \ldots, n-2$), and the distance density (the relative frequency of pairs of nodes that have a distance $l = 1, \ldots, n-2$). The comparison is done visually. The authors conclude that both the Erdős-Rényi model, and the ERGM model perform poorly in recovering the observed edgewise shared neighbors density, but do okay for degree and distance density. The authors explain that the problem is due to the transitivity in social networks and suggest that it may be avoided by building into the ERGM additional statistics relating to transitivity; e.g. triangles, k stars and/or edgewise shared neighbors.

2.2.2 AIC and BIC

We now turn to the more traditional goodness-of-fit statistics. Hunter et al. [71] suggest to use the Akaike information criterion (AIC) and the Bayes information criterion (BIC) to assess model fit in networks. For a model M with l parameters AIC and BIC are defined as

$$\begin{aligned}\mathrm{AIC}(M) &= -2(\text{maximized log-likelihood under } M) + 2 \cdot l,\\ \mathrm{BIC}(M) &= -2(\text{maximized log-likelihood under } M) + l \cdot \log(n(n-1)/2).\end{aligned}$$

Both of these measures are intended for model selection: they compare between two models instead of measuring whether a model fits well. However, if we choose the second model to be very general, AIC and BIC enable us to gain insights about the goodness-of-fit while penalizing for model complexity. Nevertheless, AIC and BIC are not applicable to many network models because we often do not know the full likelihood function. For example for the ERGM we often lack the standardizing constant κ. Furthermore, in all models based on edge dependence the A_{ij}s are not independent and thus the likelihood does not factorize.

2.2.3 Bayes factor

Using the *Bayes factor*, Latouche et al. [83] compare two logistic models for binary networks: one using only the covariates as explanatory variables (M_0), with one using the covariates plus a latent stochastic block model (M_1). The later allowing for more model complexity. The Bayes factor B is defined as

$$\mathrm{B}(M_0, M_1) = \frac{P(\boldsymbol{A}|M_0)}{P(\boldsymbol{A}|M_1)};$$

where both likelihoods $P(\boldsymbol{A}|M_0), P(\boldsymbol{A}|M_1)$ are approximated using a variational Bayes approach. As for AIC and BIC above, the Bayes factor is a relative model comparison, that only indicates which model fits better. Thus, its interpretation as a goodness-of-fit test is strongly limited by the choice of alternative. The model M_1 reflects the hypothesis that including the covariates and a latent community structure would cover all possible structure in the data; ignoring effects, for instance, like degree heterogeneity.

2.2.4 Likelihood approach

In [67], Hoff et al. answer the model fitting question by comparing maximized log-likelihoods and the number of model parameters; without directly running likelihood ratio tests. The main purpose of this paper is to introduce the latent space model as mentioned in Section 1.4.1. However, the authors compare the new latent space models for a varying number of covariates and dimensions (of the latent space), as well as, with a classic stochastic blockmodel. The authors use the maximized log-likelihoods and the number of parameters to compare the models; and mention the log-likelihood of the saturated model (where all predictions match the observations) for base-line comparison. Although considering all the ingredients, the authors do not explicitly run a likelihood ratio test since the asymptotic properties of methods where the number of parameters grows with the network size are largely unknown.

2.3 Comparison between observed networks

We here discuss two popular methods to compare observed networks. On the one hand motifs, in particular motif counts, characterize many network properties and may therefore be used to compare networks [13, 15]. For instance, an increased number of triangles is an indicator for transitivity, and a typical property for social networks (as mentioned in Section 1.3.1). For some exponential random graph models motifs are even the sufficient statistics (e.g. Markov graphs [54]). On the other hand, testing directly whether two networks are generated from the same model instead of relying on summary statistics might lead to more reliable results.

2.3.1 Motifs

Bickel et al. [15] show that motif counts enable us to fit graphons through a method of moments approach under some assumptions. The authors use motif counts to estimate the moments (the theoretical frequencies of occurrences of motifs). They derive consistency results and a central

limit theorem for acyclic motifs; both times assuming a graphon model and that the average expected degree is $o(1)$.

To identify motifs that are typical for networks of the same domain, Milo et al. [100] introduce a hypothesis test to analyze whether in a given network the number of motif occurrences significantly exceeds the expected number under a degree-based model. Applying their method to gene regulatory networks and neural networks, beside others, the authors identify those motifs that are typical for each of the domains. The aim is to extract reoccurring patterns in networks to improve our understanding of the domain specific network behavior.

To detect anomalies, Coulson et al. [35] deliver a Poisson approximation for the distribution of motif counts using the Stein-Chen method; under both a stochastic blockmodel and a graphon model. The authors derive finite-sample upper bounds for the total variation distance between a Poisson distribution and the empirical distribution of motif counts; under the assumption that for the average degree $\bar{d}$ of a subgraph $H \subseteq G$ it holds that $\bar{d}(H) < \bar{d}(G)$. This method enables us to detect for an observed network discrepancies in the connecting behavior of its nodes compared to a stochastic blockmodel or a graphon model.

To directly compare networks, Ali et al. [5] introduce *Netdis*—a score for network similarity that is based on the number of motif occurrences in local neighborhoods. To be more precise, it counts the number of m-motif occurrences in l-step neighborhoods. For $m = 3$ and $l = 2$, for instance, there are two motifs: 2-stars and triangles, and we count how many of these occur in a subnetwork of nodes surrounding node i with maximum distance two (for all nodes i). The parameters m and l need to be chosen according to the application; e.g., for a protein interaction network, the authors analyze motifs with $m = 3, 4, 5$ nodes in two-step neighborhoods ($l = 2$). To then compare two networks, we compute Netdis—a score based on all centered motif counts—to contrast the networks by their local structure. A detailed discussion of Netdis and a comparison to similar methods can be found in [150].

One of the main problems of counting motif occurrences is that it is computationally expensive, as it is polynomial in the number of nodes [13, 150]. In [6], Ali et al. address this issue by introducing a sub-sampling procedure based on neighborhoods and derive theoretical results that justify comparing networks using the Netdis statistic on a sample of similar-sized neighborhoods. They demonstrate their results on empirical and synthetic datasets indicating that often 10% of the data is sufficient for optimal comparison. Thus, the authors provide a solution to avoid the expensive computation of subgraph counts tailored to a network comparison based on the Netdis statistic.

Bhattacharyya and Bickel [13] introduce a more general solution: a bootstrap subsampling to estimate the motif counts, their variation across the network, and their approximate distributions in general. All results are derived under the assumption of the graphon model and under several sampling schemes. For instance, the authors subsample m nodes without replacement and compute the motif counts on the induced subgraphs. After repeating this procedure B times, the theoretical frequency of motif counts is estimated by the average over all bootstrap samples. The authors show for acyclic motifs that the estimator is unbiased; that its variance scales at most linearly in the density of the network and that the scaled absolute error converges in probability to 0 as B, n, and the number of nodes in the bootstrap sample approach infinity.

2.3.2 Test for an agreement in the generating model

Tang et al. [140] introduce a two-sample hypothesis test to analyze whether two networks on the same node set are generated from the same latent positions, or scaled or diagonal transformations of one another. The authors assume the networks to be finite-dimensional random dot product graphs with fixed but unknown latent positions and the nodes of the two networks to have a known correspondence (a bijective map from one node set to the other). The authors first use the spectral embedding of the adjacency matrix to estimate the latent positions in both networks as described in [138]. Then, they compute a test statistic based on the two estimated positions and show that the probability for their test to reject given the two true latent positions are equal (up to orthogonal transformation) is upper bounded by a significance level. Furthermore, the authors derive assumptions under which the power of the test (the probability to reject if the true latent positions are different) converges to one. The assumptions ensure beside others that the network is not too sparse and the gap of the eigenvalues is sufficiently large. The authors demonstrate their results on simulated and observed data.

2.4 Clustering in networks

As mentioned in the motivation, across sciences we observe that networks divide naturally into communities. There are numerous approaches for network community detection that aim at identifying a single "best" community assignment. All of these optimize a measure for the quality of a group assignment differing in their motivation: there are heuristic and model-based approaches. Furthermore, all approaches mentioned below assume the number of communities K to be known (unless otherwise specified). We introduce both kinds of community detection

methods and finish with a discussion of how to estimate K.

The main difference between our work and the work presented in this section is that we step away from the idea of a single "best" community assignment. We instead acknowledge that each of several community assignments may describe different aspects of a network's structure. We define these community assignments based on covariates; adding interpretability to the community structure. Hence, instead of using the covariates to improve the community detection as done by some of the approaches below, we derive methodology to evaluate *observed* community structure implied by the covariates themselves.

2.4.1 Model-based community detection

In this subsection, we present a selection of model-based methods for community detection to illustrate different approaches how to use the likelihood as optimization criterion. Note that each of the models that support community structure described in Section 1.4 may be utilized for community detection even if not listed below. Given the interest in covariates for this thesis, we sort the methods into those solely based on the network and those that incorporate covariates.

Community detection solely based on the network

Snijders and Nowicki [135] infer the community membership using *Bayesian estimation* of a *stochastic blockmodel* with two groups for simple networks. The authors estimate the group membership $\boldsymbol{g}$ based on its posterior distribution conditioned on the observed network $\boldsymbol{A}$. With $\omega_{lk} = \Pr(A_{ij} = 1 | g(i) = l, g(j) = k)$ and $\eta_k = \Pr(g(i) = k)$ for all i, j, we obtain

$$\Pr(\boldsymbol{g}|\boldsymbol{A}) = \int f(\boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\eta}) \, d\boldsymbol{\omega} d\boldsymbol{\eta}.$$

Gibbs sampling is used to derive $f(\boldsymbol{g}, \boldsymbol{\omega}, \boldsymbol{\eta})$. The parameters $\boldsymbol{\omega}$ and $\boldsymbol{\eta}$ are unidentifiable because the distribution of $\boldsymbol{A}$ is invariant under label permutation. However, it can be solved by introducing a restriction on the parameter space. The authors show that the community membership is recovered correctly with high probability for dense networks $\big(\mathbb{E}\, N = \Omega\big(n^2\big)\big)$ under a stochastic blockmodel. We will refer to this property henceforth as strong consistency. For an explanation of "with high probability" see Appendix A.1 Definition 8. Nowicki and Snijders [115] generalize this approach for directed, weighted networks allowing for more than two communities.

Following a frequentist approach, Bickel and Chen [14] infer the community membership by maximizing the *profile likelihood* of a *stochastic blockmodel*. The authors assume the

number of communities K to be fixed (i.e. independent of n). Let n_{lk} denote the number of edges connecting groups l and k and n_l the number of nodes in group l. The authors follow a two-step procedure. First, when conditioning on a group assignment $\boldsymbol{g}$, the maximum of the conditional log-likelihood is obtained at $\hat{\omega}_{lk} = n_{lk}/n_l n_k$—the fraction of observed edges connecting groups l and k. Second, they estimate $\boldsymbol{g}$ as the argument that maximizes the profile likelihood:

$$l(\boldsymbol{g}|\boldsymbol{A}, \hat{\boldsymbol{\omega}}) = \frac{1}{2} \sum_{1 \leq l,k \leq K} \left[n_{lk} \log\left(\frac{n_{lk}}{n_l n_k}\right) + (n_l n_k - n_{lk}) \log\left(1 - \frac{n_{lk}}{n_l n_k}\right) \right].$$

The authors show strong consistency of the estimator for the community assignment under the stochastic blockmodel for sparse networks ($\mathbb{E}\, N = \omega(n \log n)$). Choi et al. [29] generalize the maximum profile likelihood fitting by allowing the number of communities to grow; i.e., $K = \mathcal{O}(\sqrt{n})$, and then show that the relative number of misclassified nodes converges in probability to 0 under the stochastic blockmodel for sparse networks $\left(\mathbb{E}\, N = \omega\left(n[\log n]^{3+\epsilon}\right), \epsilon > 0\right)$. Thus, Choi et al. generalize the setting in [14] but show a weaker property. We will refer to this property as weak consistency.

Assuming we observe the group labels for all but a few nodes in a simple network, Sussman et al. [138] suggest a *semi-supervised clustering* approach based on the *random dot product graph*. The authors suggest predicting the missing community assignments by applying a K-nearest neighbor clustering to the estimated latent positions $\widehat{\boldsymbol{Z}}$. Recall from Section 1.4.1, to estimate the latent positions the authors firstly compute the eigen-decomposition $\left(\boldsymbol{A}\boldsymbol{A}^T\right)^{1/2} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{U}^T$. Secondly, they define a diagonal matrix $\boldsymbol{S}_{[d]} \in \mathbb{R}^{d \times d}$ of the d largest eigenvalues and define $\boldsymbol{U}_{[d]} \in \mathbb{R}^{n \times d}$ as the matrix with the corresponding eigenvectors; to then

$$\widehat{\boldsymbol{Z}} = \boldsymbol{U}_{[d]} \boldsymbol{S}_{[d]}^{1/2}.$$

The authors cluster the rows of $\widehat{\boldsymbol{Z}}$ using a K-nearest neighbor approach and show that this method leads to a weakly consistent estimator of the community assignments under a random dot product graph. This approach is closely related to a heuristic clustering method called spectral clustering (see Section 2.4.2).

Community detection that incorporates covariate information

Illustrating the growing interest in the relation between covariates and community structure, many methods for community detection, including those based on the models in Section 1.4, have recently been extended to improve community detection using the information captured

in covariates [50, 63, 110]. We now present an illustrative selection of model-based methods rather than an exhaustive list.

As mentioned in Subsection 1.4.1, Handcock et al. [63] extend the *latent space model* that always incorporates covariates to include community structure by modeling the latent positions as a mixture of Normal random vectors. To be more precise, the authors assume that

$$\begin{aligned}\operatorname{logit}\Pr(A_{ij}=1|\boldsymbol{z}_i,\boldsymbol{z}_j,\boldsymbol{x}_{ij},\boldsymbol{\beta}) &= \boldsymbol{\beta}_0^T\boldsymbol{x}_{ij}-\beta_1|\boldsymbol{z}_i-\boldsymbol{z}_j|,\\ \boldsymbol{z}_i &\overset{iid}{\sim} \sum_{k=1}^{K}\lambda_k \operatorname{Normal}\big(\boldsymbol{\mu}_k,\sigma_k^2\boldsymbol{I}_d\big),\end{aligned}$$

where λ_k is the probability for a node to belong to community k, and thus $\sum_{k=1}^{K}\lambda_k = 1$. $\boldsymbol{I_d}$ denotes the $(d\times d)$-identity matrix and $\boldsymbol{\mu}_k$ identifies the center of the k-th community in the social space. The authors fit the model in a two stage procedure: first estimating the latent positions as in the (non-clustering) latent space model (see Section 1.4.1) and second applying a maximum likelihood fitting to the normal mixture model conditioned on the estimated latent positions $\hat{\boldsymbol{z}}_1,\ldots,\hat{\boldsymbol{z}}_n$. Since the likelihood is not strictly concave, care must be taken when initializing algorithms like the expectation maximization. The authors suggest selecting the number of communities K by comparing several values using a Bayes information criterion. However, the authors point out that for the examples in the paper this method of choosing K performs poorly.

Newman and Clauset [110] introduce a Bayesian community detection approach based on a *degree-corrected stochastic blockmodel* where the prior probability of the community assignment $\boldsymbol{g}$ depends on a node's covariates. The authors estimate the node-specific parameters to match the observed degree, and fit the remaining model using a maximum likelihood approach realized by an expectation-maximization algorithm. For the community detection the authors deliver the marginal posterior probabilities of the group assignments. In addition, the authors estimate the prior probability of the node to belong to a group given its covariates, thereby quantifying the correlation between the covariate and the optimal community structure.

2.4.2 Heuristic community detection

In this work, we present three popular heuristically motivated approaches [109]. In this subsection, we start with two approaches that are based on either eigen or singular value decomposition. These methods are less computationally demanding than methods based on likelihood maximization (see Sections 2.4.1) because there are fast algorithms available for eigen and sin-

gular value decomposition, especially for sparse matrices. The third heuristically motivated approach is called modularity maximization and since it is crucial to the thesis, it will be discussed in more detail in Section 2.5.

Spectral clustering [43, 51] assigns communities in an undirected, weighted network such that the least number of edges (or edges of low weight) is connecting the communities [92]. This optimization criterion is often referred to as the minimum-cut problem. Formally, for K communities $D_1, \ldots, D_K$ we measure a cut distance as

$$\mathrm{NCut}(D_1, \ldots, D_K) = \frac{1}{2} \sum_{k=1}^{K} \frac{\#\text{ edges between } D_k \text{ and its complement } V \backslash D_k}{\sum_{i \in D_k} d_i}.$$

Solving this minimum-cut problem is NP-hard [92]. The main insight of spectral clustering is that the discrete optimization problem can be well approximated by a continuous relaxation. The solution of the continuous optimization problem can be derived from clustering the rows of an eigenvector matrix of the *graph Laplacian* $\boldsymbol{L}$:

$$\boldsymbol{L} = \boldsymbol{I} - \boldsymbol{D}^{-1}\boldsymbol{A},$$

where $\boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_n)$ denotes the diagonal matrix of the degrees. To be more precise, since $\boldsymbol{L}$ is positive semidefinite, it has n non-negative real-valued eigenvalues. We define a matrix $\boldsymbol{U} \in \mathbb{R}^{n \times K}$ of the eigenvectors $\boldsymbol{u_1}, \ldots, \boldsymbol{u_K}$ corresponding to the K smallest eigenvalues of $\boldsymbol{L}$. Denote $\boldsymbol{y_i} \in \mathbb{R}^K$ the i-th row of $\boldsymbol{U}$. Cluster the points $\boldsymbol{y_1}, \ldots, \boldsymbol{y_n}$ using K-means algorithm into clusters $C_1, \ldots, C_K$. Then, the community structure $\boldsymbol{g} = (g(i) = k \text{ if } \boldsymbol{y_i} \in C_k, i = 1, \ldots, n)$ has a cut distance close to the global minimum [92]. For a detailed motivation of spectral clustering see [92].

Rohe et al. [126] show that under a stochastic blockmodel for simple networks spectral clustering leads to weakly *consistent* estimators for the community membership: the fraction of misclassified nodes converges to 0 in n; while allowing the number of communities to grow in n. The two main assumptions are that the network is not too sparse ($\min_i \mathbb{E}\, d_i$ grows almost linearly in n) and that the smallest non-zero eigenvalue of the "population" graph Laplacian $\boldsymbol{L}$ shrinks slowly enough (leading to a small eigengap—the difference between the leading eigenvalues and the others). Lei et al. [87] generalize this result for networks that are more sparse ($\max_i \mathbb{E}\, d_i \leq \alpha_n \leq \log n$) and under the degree-corrected stochastic blockmodel.

Many *modifications* of the graph Laplacian have been suggested: Sarkar and Bickel [129] demonstrate that if we normalize the graph Laplacian; i.e., $\boldsymbol{L} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, it reduces the variance of the estimators of the community assignment under the stochastic blockmodel

asymptotically; compared to the unnormalized spectral clustering where $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$. To reduce the sensitivity of spectral clustering to sparsity, Amini et al. [7] suggest regularizing the graph Laplacian by adding a constant perturbation of τ/n to every entry of $\boldsymbol{A}$. Joseph and Yu [74] show that the regularized graph Laplacian concentrates for large τ: improving the performance of the k-means of the rows of the eigenvector matrix; while the eigengap decreases: complicating the estimation of the eigenvectors. Thus, the optimal τ balances these two effects. The authors show that the regularization allows us to weaken the assumption on sparsity and prove that regularized spectral clustering is consistent.

Sussman et al. [137] suggest a community detection approach for *directed* networks that works with a low rank approximation of the adjacency matrix itself. Let $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ be the *singular value decomposition* of $\boldsymbol{A}$ and denote $\boldsymbol{U}_{[d]}$ and $\boldsymbol{V}_{[d]}$ the first d columns of the orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$; and $\boldsymbol{\Sigma}_{[d]}$ the diagonal matrix with the d largest singular values of $\boldsymbol{A}$ on the diagonal. Sussman et al. suggest approximating the adjacency matrix with $\boldsymbol{X}\boldsymbol{Y}^T$ where $\boldsymbol{X} = \boldsymbol{U}_{[d]}\boldsymbol{\Sigma}_{[d]}^{1/2}$ and $\boldsymbol{Y} = \boldsymbol{V}_{[d]}\boldsymbol{\Sigma}_{[d]}^{1/2}$. The authors then cluster the rows $\boldsymbol{w}_i \in \mathbb{R}^{2d}$ of the matrix $\boldsymbol{W} = \left(\boldsymbol{U}_{[d]}, \boldsymbol{V}_{[d]}\right) \in \mathbb{R}^{n \times 2d}$ by optimizing the following criterion

$$\left(\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{g}}\right) = \text{argmax}_{\boldsymbol{\psi}, \boldsymbol{g}} \sum_{i=1}^{n} \left\| \boldsymbol{w}_i - \boldsymbol{\psi}_{\boldsymbol{g(i)}} \right\|_2^2$$

where $\boldsymbol{\psi_k} \in \mathbb{R}^{2d}$ is the centroid of block k, for $k = 1, \ldots, K$. The authors show that the number of misclassified nodes goes to 0 with high probability under the stochastic blockmodel.

2.4.3 Identification of the number of communities

Many of the methods above assume the number of communities K to be known while in practice, that is hardly ever the case. We now discuss three methods addressing how to infer this number from data, selected to demonstrate the three main approaches.

Bickel and Sarkar [16] introduce a *spectral method* to simultaneously identify the number of communities K and detect the optimal community assignment. The authors suggest a recursive bipartitioning algorithm that as a stopping criterion, runs at each step a hypothesis test between an Erdős-Rényi graph (null) and a stochastic blockmodel (alternative). Under the null, let us assume a dense Erdős-Rényi graph $G(n, p)$ (i.e. $\mathbb{E}\, N = \Theta(n^2)$). Compute $\boldsymbol{A}^*$ by centering and scaling of the adjacency matrix $\boldsymbol{A}$ of a simple network with $\hat{p} = N/(n(n-1))$:

$$A_{ij}^* = \begin{cases} \frac{A_{ij} - \hat{p}}{\sqrt{(n-1)\hat{p}(1-\hat{p})}} & \text{for } i \neq j \\ 0, & \text{for } i = j. \end{cases}$$

Then, Bickel and Sarkar show for the largest eigenvalue λ_1 of A^* that it holds that

$$\lambda_1^* = n^{2/3}[\lambda_1(\boldsymbol{A}^*) - 2] \xrightarrow{d} \text{Tracy-Widom}(1).$$

In contrast, under the alternative of a stochastic blockmodel with $K > 1$ and more edges within than between communities, the authors show that $\lambda_1^* \to \infty$. Hence, we can compute a p-value $\Pr(X \geq \lambda_1^*)$ under the null hypothesis of a Tracy-Widom distribution while being protected against the alternative hypothesis. In each step, the authors run the test; if it rejects the null, the network is partitioned into two subgraphs; and one repeats the procedure for each of the subgraphs. As a consequence, the authors deliver an algorithm that detects hierarchical community structure and automatically determines the number of communities. Extending these results, Lei [86] conducts a model selection between stochastic blockmodels with different number of blocks K without running a recursive procedure. The author derives the asymptotic distribution of the test statistic under a stochastic blockmodel (null), computes an upper bound for the eigenvalues under an alternative model with larger K, and shows strong consistency of the estimator for K. The author assumes the communities to be distinguishable (loosely speaking) and $K = \mathcal{O}\big(n^{1/6+\tau}\big)$, for some $\tau > 0$.

Newman and Reinert [113] estimate the number of communities K as the mode of the *posterior distribution* $\Pr(K|\boldsymbol{A})$ under the assumption of a stochastic blockmodel. For ease in mathematical difficulty, the authors focus on multi-edge networks with self-loops since it enables them to model the edges as Poisson distributed. Recall that n_{lk} denotes the number of edges connecting groups l and k, n_k the number of nodes in group k, and ω_{lk} the probability for nodes of communities l and k to connect. Further denote γ_k the probability for a node to be in community k. Then, we obtain the likelihood

$$\begin{aligned}\Pr(\boldsymbol{A}, \boldsymbol{g}|\boldsymbol{\omega}, \boldsymbol{\gamma}, K) &= \Pr(\boldsymbol{g}|\boldsymbol{\gamma}, K)\Pr(\boldsymbol{A}|\boldsymbol{\omega}, \boldsymbol{g}) \\ &= \prod_{k=1}^{K} \gamma_{kk}^{n_{kk}} \prod_{k} \omega_{kk}^{n_{kk}} \exp\big(-n_k^2 \omega_k k/2\big) \prod_{k<l} \omega_{kl}^{n_{kl}} \exp(-n_k n_l \omega_l k).\end{aligned}$$

The authors assume a uniform (least informative) prior on K, and γ with the additional condition that $\sum_{k=1}^{K} \gamma_k = 1$. Following an empirical Bayes approach, the authors set the mean of the edge probabilities $\omega_{lk}, l, k = 1, \ldots, K$ equal to the fraction of the observed edges while using an exponential distribution as a prior. Under these assumptions, the authors derive $\Pr(K, \boldsymbol{g}|\boldsymbol{A})$. Using Markov chain Monte Carlo importance sampling, they obtain $\Pr(K|\boldsymbol{A})$ as the histogram of the Monte Carlo samples. The authors provide a generalization for the degree-corrected

stochastic blockmodel. The authors demonstrate the performance of their approach using simulated and observed data.

Wang and Bickel [143] suggest a *likelihood ratio test* for selecting K based on the stochastic blockmodel for simple networks. Denote f the likelihood under the stochastic blockmodel and K_0 the true number of communities. The authors define a quality function

$$\beta\big(K'\big) = \left(\sup_{\boldsymbol{\omega},\boldsymbol{\pi}} \log \sum_{\boldsymbol{g}\in\{1,\dots K'\}^n} f(\boldsymbol{g},\boldsymbol{A}|\boldsymbol{\omega},\boldsymbol{\pi}) - N_{K'}B_n \right)$$

where $N_{K'}$ is a strictly increasing sequence in K' and

$$\begin{aligned} B_n &= o(n\,\mathbb{E}\,N) && \text{for } K' < K_0, \\ B_n &: B_n n^{-1/2}(\mathbb{E}\,N)^{-1/2} \to \infty && \text{for } K' > K_0. \end{aligned}$$

The authors suggest selecting K such that

$$K = \operatorname{argmax}_{1\le K'\le n} \beta\big(K'\big)$$

and show that this model selection approach is consistent:

$$\Pr\big(\beta\big(K'\big) < \beta(K_0)\big) \to 1.$$

The main step in the proof shows that under some regularity conditions, and for $K' \neq K_0$ it holds that the likelihood ratio test statistic

$$L_{K_0,K'} = \log \frac{\sup_{\boldsymbol{\omega},\boldsymbol{\pi}} \sum_{\boldsymbol{g}\in\{1,\dots K'\}^n} f(\boldsymbol{g},\boldsymbol{A}|\boldsymbol{\omega},\boldsymbol{\pi})}{\sup_{\boldsymbol{\omega},\boldsymbol{\pi}} \sum_{\boldsymbol{g}\in\{1,\dots K_0\}^n} f(\boldsymbol{g},\boldsymbol{A}|\boldsymbol{\omega},\boldsymbol{\pi})}$$

is asymptotically well behaved:

$$\begin{aligned} \frac{n^{-3/2}L_{K_0,K'} - \sqrt{n}\mu}{\sigma^2} &\xrightarrow{d} \text{Normal}(0,1) && \text{for } K' = K_0 - 1, \\ L_{K_0,K'} &= \mathcal{O}_P\Big(n^{1/2}\,\mathbb{E}\,N\Big) && \text{for } K' > K_0. \end{aligned}$$

The authors provide μ and σ for different sparsity regimes.

2.5 Clustering in networks via modularity

In this section, we present a heuristic community detection method called *modularity maximization* [111] that with over 7000 citations (Google scholar, August 2016) is a popular community detection method. Modularity maximization is discussed here in more detail than all the other

community detection approaches above because the original work of this thesis builds up on the concept of modularity. This section is organized as follows. We start by introducing modularity maximization conceptually as well as algorithmically. We then discuss its advantageous and disadvantageous properties; to conclude by presenting results that relate the original modularity maximization to our work.

2.5.1 Introduction

Modularity maximization is similarly to spectral clustering a community detection method motivated by the minimum-cut problem. Its concept is based on maximizing an intuitive and practically effective measure of community structure called modularity. For a given $\boldsymbol{g}$, modularity computes the difference between the observed edges A_{ij} within communities, and the estimated expected number of edges $d_i d_j / \|\boldsymbol{d}\|_1$, in the absence of community structure. To be more precise, modularity is defined as follows.

Definition 1 (Modularity [107])**.** *The Newman–Girvan modularity is given by*

$$\widehat{Q}(\boldsymbol{A}, \boldsymbol{g}) = \begin{cases} \sum_{j=1}^{n} \sum_{i<j} \left[A_{ij} - \frac{d_i d_j}{\|\boldsymbol{d}\|_1} \right] \delta_{g(i)=g(j)}, & \text{if } \|\boldsymbol{d}\|_1 \neq 0 \\ 0, & \text{otherwise;} \end{cases} \tag{2.1}$$

where $\delta_{g(i)=g(j)} = 1$ *when nodes* i *and* j *are in the same group, and* 0 *otherwise.*

Modularity is typically standardized (i.e., divided) by $\|\boldsymbol{d}\|_1$ and then takes values in $(-1, 1)$. A positive value indicates assortative community structure since we observe more edges within communities than expected in the absence of community structure. A negative value indicates a disassortative community structure where there is a tendency for connections to be between groups rather than connecting nodes of the same group. A value close to 0 indicates that there is no clear tendency for the community structure to be either assortative or disassortative. For instance, a mixture of strongly assortative and disassortative communities can lead to a small modularity value since the summands may cancel out. For notational convenience we write modularity $\widehat{Q}(\boldsymbol{A}, \boldsymbol{g})$ henceforth as $\widehat{Q}$.

Algorithmically, Newman and Girvan [111] introduce a hierarchical clustering to find the community assignment that maximizes modularity. An exact solution to this problem is NP-hard. The authors instead partition the network into communities by iteratively removing the edge with the highest shortest-path betweenness score: a count of the shortest paths between all pairs of nodes that pass through the edge under study. This leads to a selection of community assignments differing beside others in the number of communities K. The authors then introduce

modularity as a quality measure for community structure and select the community assignment with the highest modularity value. Much work has been conducted to improve the modularity maximization algorithm [34, 106, 107], with one of the fastest being the Louvain algorithm [19] that is built into a number of software packages for network analysis. The original algorithms in [107, 111] on modularity maximization are designed for simple networks but have been extended since to directed networks, multi-layer networks, and weighted networks [88, 102, 105].

Zhang et al. [153] extend modularity to incorporate covariates. The authors alter modularity by introducing an additional weight on the edges, where high weights indicate similar covariate values of the end nodes. The influence of the covariates may vary across communities and between covariates. The authors then optimize the joint criterion on both the community assignment and the impact of the covariates; by fixing one and optimizing for the other alternately. The authors show that this community detection approach is strongly consistent under the stochastic blockmodel in the sparse case regime where $\mathbb{E}\, N = \omega(n \log n)$.

2.5.2 Properties

We now describe the advantages and disadvantages of modularity.

Modularity maximization has proven to be practically useful across sciences [60, 94, 127]. The algorithm simultaneously estimates the community assignment $\boldsymbol{g}$ and the number of communities K [111]; while for spectral clustering and most of the maximum likelihood approaches above we need to choose K in advance. Furthermore, Zhao et al. [154] show that while the likelihood-based criteria (e.g., [75]) are theoretically preferable (fewer assumptions), the algorithm for modularity that relates it to an eigenvalue problem [107] beats computationally many of the likelihood-based solutions. As for many other approaches, modularity maximization leads to a consistent estimator for the community assignment under the degree-corrected stochastic blockmodel; assuming loosely speaking, that the edges within communities are more likely than between communities and that the network is not too sparse [154]. In fact, if it holds for the network that $\mathbb{E}\, N = \omega(n)$ or $\mathbb{E}\, N = \omega(n \log n)$, then it follows weak or strong consistency, respectively [154]. Note that Bickel and Chen [14] were the first to discuss consistency for modularity but under the less realistic stochastic block model.

A disadvantage of modularity as pointed out by Newman and Reinert [113] is that it suffers from only being heuristically motivated. For instance, it is not objective as a measure of the strength of community structure. Newman and Girvan [107] propose based on experience with real-world networks that a value above 0.3 indicates a strong community structure; implying

that the modularity values were comparable between networks. However, Good et al. [61] show that the modularity values depend on the size of the network and the number of communities. As a first step to address this problem, McDiarmid and Skerman [97] give an upper bound on the modularity value for two special cases: a random r-regular graph (each node has r neighbors, and the connection is made at random) and for subgraphs of integer lattices. Furthermore, modularity admits a resolution limit: a preference for communities that are above a minimum size relative to the total number of nodes and the interconnectedness of the communities [52]. For instance, it may assign a smaller value to two loosely connected cliques than to their union if the cliques are sufficiently small.

2.5.3 Related work

The work presented in this section so far discusses modularity as a community detection method. In contrast, our work extends its applicability to objectively quantify the strength of observed community structure; thereby stepping away from the idea of a single "best" community assignment in favor of a covariate-based community structure (see Chapter 4). In this subsection, we present the work closest to ours that, to some extent, we build up on. It is sorted chronologically.

Arias-Castro and Verzelen [8] utilize a simplified version of modularity to detect the presence of network community structure. They test the null model of an Erdős-Rényi graph $G(n, p_0)$ on n nodes with success probability p_0 against a nested alternative of an Erdős-Rényi subgraph $G(n_1, p_1)$ within the Erdős-Rényi graph $G(n, p_0)$ with $n_1 \leq n$. The authors show that the test is asymptotically powerful (Type I and Type II errors converge to 0 in n) if the Kullback-Leibler divergence between p_0 and p_1 is large enough relative to n_1. This work is a first step to build a theoretical foundation for modularity. However, the change from the degree-based model as implied by Newman and Girvan's modularity to an Erdős-Rényi graph leads to a drastic simplification of the problem.

Volfovsky and Hoff [142] introduce a hypothesis test for the independence of nodes in a network based on row and column correlations in the adjacency matrix. The authors consider directed, weighted networks and model the adjacency matrix as

$$\operatorname{vec} \boldsymbol{A} \sim \operatorname{Normal}(\mathbf{0}, \boldsymbol{\Sigma_c} \otimes \boldsymbol{\Sigma_r}),$$

where the operator vec vectorizes the adjacency matrix; Σ_c and Σ_r represent the covariances between nodes as senders and as receivers, respectively; and $\otimes$ denotes a Kronecker product.

The authors test a null of independent edges against an alternative of correlated edges using a likelihood ratio test. Using Monte-Carlo procedures, the authors approximate the distribution of the test statistic both under the null and alternative model. In contrast to our results, this work does not address covariates or community structure and the statistical guarantees are based on simulations rather than asymptotic theory.

After the posting of [56] describing the work of this thesis, Newman [109] shows that optimizing a generalized modularity (see Eq. (2.3), [123]) is equivalent to maximizing the conditional likelihood of the degree-corrected stochastic blockmodel when the true probabilities $\boldsymbol{\omega}$ are known; and under the condition that $\hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}$. Newman assumes a degree-corrected stochastic blockmodel (see Section 1.4.1) where $\omega_{kk} = \omega_{in}$ for all k; and $\omega_{kl} = \omega_{out}$ for all $l \neq k$. The author estimates the node-specific parameters as the scaled degrees: $\hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}$ and substitutes $\hat{\pi}_i$ for π_i for all i. As in [75], the author models the edges as Poisson distributed; assuming a multi-edge network with self-loops to ease the computations. Under these assumptions, we may write the conditional log-likelihood as

$$\log \Pr\left(\boldsymbol{A}\middle|\boldsymbol{\omega}, \boldsymbol{g}, \hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}, \forall i\right) = B\sum_{i<j}\left(A_{ij} - \gamma\frac{d_i d_j}{\|\boldsymbol{d}\|_1}\right)\delta_{g(i)=g(j)} + C, \tag{2.2}$$

where B, C and γ are constants that only depend on $\boldsymbol{\omega}$ but not on $\boldsymbol{g}$ directly. Under the assumption that $\boldsymbol{\omega}$ is given, and apart from constants that only depend on $\boldsymbol{\omega}$ the conditional log-likelihood of this degree-corrected stochastic blockmodel is equal to a generalized modularity (introduced in [123] to circumvent the resolution limit):

$$\Pr(\boldsymbol{A}|\boldsymbol{\omega}, \boldsymbol{g}) = \sum_{i<j}\left(A_{ij} - \gamma\frac{d_i d_j}{\|\boldsymbol{d}\|_1}\right)\delta_{g(i)=g(j)}. \tag{2.3}$$

Note that in practice to fit the maximum likelihood for the degree-corrected stochastic blockmodel conditioned on the $\hat{\pi}_i$s, we need to estimate $\boldsymbol{\omega}$ and its estimator strongly depends on the estimator for $\boldsymbol{g}$. In turn, when estimating $\boldsymbol{\omega}$, the constants B, C and γ in Eq. (2.2) depend on $\boldsymbol{g}$ and are not negligible. However, Newman's work shows us that community detection using modularity maximization is strongly related to the maximum likelihood approach for the degree-corrected stochastic blockmodel.

Chapter 3

Nonparametric family of degree-based models

In this chapter, we introduce a nonparametric family of degree-based models, show how to fit these models, and deliver statistical guarantees for their estimators. Network models that assume the propensity of all connections to be governed solely by a single parameter per node are frequently used in practice, because they are easy to understand and analyze. A node's parameter may be interpreted as a measure of its centrality, and fitting these models therefore improves our understanding of the relations between the nodes. As we have seen in Section 1.4 in the introduction, the first of these models was introduced by Chung and Lu [30], and henceforth has been used to model binary networks.

In outline, we extend the definition by Chung and Lu [30] to a broad class of network models, covering weighted, multi-edge, and power-law networks (Section 3.1). For simple networks, Perry and Wolfe [119] introduce estimators for the model parameters and show that these are near-maximum likelihood estimators in the sparse graph regime. In Section 3.2, we derive the asymptotic distribution of these estimators in our more general framework, and provide confidence intervals to quantify their uncertainty. The former results extend work in [116]. In Section 3.3, we establish the implications of these results on estimating the expectation of an edge $\mathbb{E}\, A_{ij}$; showing weak consistency and convergence in distribution for the resulting estimators. After illustrating our results using a simulation study (Section 3.4), we conclude the chapter with a discussion of the practical implications of the theory derived here (Section 3.5).

Not only enable us the results in Section 3.2 to decide whether two nodes differ significantly in their centrality while controlling for a Type I error, they also give rise to a well performing estimator for an edge expectation (see Section 3.3). As we will explain in Chapter 4, the family of degree-based models in addition plays an important role for modularity as a null model that indicates a lack of community structure. A degree-based model cannot readily describe networks whose aggregate links behave in a block-like manner.

3.1 Definition of a nonparametric family of degree-based models

Definition 2 (The family of degree-based models)**.** *Consider an undirected, random graph on n nodes without self-loops. We model its (possibly weighted) edges $A_{ij} \geq 0$ as independent random variables with expectations given by the product of node-specific parameters $\pi_1, \pi_2, \ldots, \pi_n > 0$:*

$$\mathbb{E}\, A_{ij} = \pi_i \pi_j, \quad 1 \leq i < j \leq n.$$

Furthermore, considering a sequence of such networks as n grows, we assume they are well behaved asymptotically:

1. *No single node dominates the network:* $\max_i \pi_i / \bar{\pi} = \mathcal{O}(1)$*, with* $\bar{\pi} = \frac{1}{n}\sum_{l=1}^{n} \pi_l$*;*
2. *The network is not too sparse:* $\min_i \pi_i = \omega(1/\sqrt{n})$*;*
3. *The expectation of each edge $\mathbb{E}\, A_{ij}$ does not diverge too quickly as n grows:* $\max_i \pi_i = o(\sqrt{n})$*;*
4. *The variance of each edge does not vary too much from its expectation:* $\forall i, j$: $\operatorname{Var} A_{ij} / \mathbb{E}\, A_{ij} = \Theta(1)$*; and*
5. *The skewness of each edge A_{ij} is controlled:* $\forall i, j$: $\mathbb{E}\Big[(A_{ij} - \mathbb{E}\, A_{ij})^3\Big] \Big/ \operatorname{Var}(A_{ij}) = \mathcal{O}(1)$.

We make no further assumptions on the distribution of A_{ij}, and so our results apply in many settings, including weighted networks and those with multiple edges. Assumptions 1–3 are structural: the first excludes star-like networks; the second ensures that the network is not too sparse $\big(\Rightarrow \mathbb{E}\, N = \omega\big(n^{3/2}\big)\big)$; and the third controls the growth of $\mathbb{E}\, A_{ij}$ with n in the weighted or multi-edge setting. Assumptions 4 and 5 are technical; they exclude extreme behavior of the edge variables. For instance, both are fulfilled whenever $A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j)$ or $A_{ij} \sim \text{Poisson}(\pi_i \pi_j)$.

To characterize the family of degree-based models, we now explain how to fit these models and establish the large-sample properties of their estimators.

3.2 Properties of the estimator of a node's centrality

Each parameter π_i describes the relative centrality of node i. Thus, to fit the degree-based model of Definition 2 to a network, we estimate the parameters π_i using the node's degrees d_i

as follows [31, 116, 119]:

$$\hat{\pi}_i = \frac{d_i}{\sqrt{\|\boldsymbol{d}\|_1}}, \quad 1 \leq i \leq n. \tag{3.1}$$

The estimator $\hat{\pi}_i$ is both more natural and more computationally efficient than the corresponding maximum-likelihood estimator for π_i, which follows from the theory of generalized linear models and cannot be written explicitly in closed form. In many settings the difference between these estimators is provably small [119], and so properties of maximum likelihood estimation can also be expected to hold for Eq. (3.1).

3.2.1 A limit theorem for the estimator of a node's centrality

We show in Theorem 3.2.1 below that the estimator defined by Eq. (3.1) tends toward a Normal distribution when n is large and Definition 2 is in force. As a consequence, we may not only conclude large-sample properties for $\widehat{\mathbb{E}\, A_{ij}}$, as we will show in Corollary 3.3.2 in the next section, Theorem 3.2.1 also gives rise to a z-test. It enables us to identify pairwise differences in centrality $\pi_i, i = 1, \ldots, n$. For example, let us choose two out of 153 employees at random from an email interaction network (see Enron data in Chapter 5). We can test whether the employees 4 and 28 are equally central ($H_0 : \pi_4 = \pi_{28}$ vs. $H_1 : \pi_4 \neq \pi_{28}$). It follows from Theorem 3.2.1 that the test statistic $T = (\hat{\pi}_4 - \hat{\pi}_{28})/\sqrt{0.5(\operatorname{Var}\hat{\pi}_4 + \operatorname{Var}\hat{\pi}_{28})}$ is approximately $\mathrm{Normal}(0, 1)$ under H_0 and we can therefore compute the probability $p = 0.038$ that $|T|$ is greater than the observed value 2.08. Hence, we reject the hypothesis that employees 4 and 28 are equally central. Theorem 3.2.1 generalizes parts of the results in [116] by Olhede and Wolfe where the authors assume $\mathrm{Bernoulli}(\pi_i\pi_j)$ edges and a power law degree distribution.

Theorem 3.2.1 (Central limit theorem for Eq. (3.1))**.** *Assume the model of Definition 2 and define an estimator $\hat{\pi}_i$ of π_i as in Eq.* (3.1)*. Then as $n \to \infty$,*

$$\frac{\hat{\pi}_i - \pi_i}{\sqrt{\operatorname{Var} d_i / \mathbb{E}\|\boldsymbol{d}\|_1}} \xrightarrow{d} \mathrm{Normal}(0, 1).$$

Furthermore, $\sqrt{\operatorname{Var} d_i / \mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}(1/\sqrt{n})$, and can be consistently estimated using a plug-in estimator for $A_{ij} \sim \mathrm{Bernoulli}(\pi_i\pi_j)$ and $A_{ij} \sim \mathrm{Poisson}(\pi_i\pi_j)$ by substituting $\hat{\pi}_i$ for π_i in $\mathbb{E}\, d_i$ and $\operatorname{Var} d_i$.

Proof. The proof is a generalization of the proof of Theorem 3.2 in [116], which assumes Bernoulli edges and a power law degree distribution. First, we set the preliminaries: Since the

edges $A_{ij}, i < j$ are independent, it follows as shown in [116] that for finite n

$$\mathbb{E}\, d_i = \pi_i(\|\boldsymbol{\pi}\|_1 - \pi_i), \qquad \operatorname{Var} d_i = \sum_{i \neq j} \operatorname{Var} A_{ij}, \tag{3.2}$$

$$\mathbb{E}\|\boldsymbol{d}\|_1 = \|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2, \qquad \operatorname{Var}\|\boldsymbol{d}\|_1 = 2\sum_{i=1}^{n} \operatorname{Var} d_i, \tag{3.3}$$

$$\operatorname{cov}(d_i, d_j) = \begin{cases} \operatorname{Var} A_{ij}, & i \neq j \\ \operatorname{Var} d_i, & i = j. \end{cases} \tag{3.4}$$

We are now ready to proceed with our analysis. We write

$$\frac{\hat{\pi}_i - \pi_i}{\sqrt{\operatorname{Var} d_i / \mathbb{E}\|\boldsymbol{d}\|_1}} = \left[\underbrace{\frac{d_i - \mathbb{E}\, d_i}{\sqrt{\operatorname{Var} d_i}}}_{T_1} + \underbrace{\frac{\mathbb{E}\, d_i - \pi_i\sqrt{\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}}}_{T_2} \right] \underbrace{\sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1}}}_{T_3}. \tag{3.5}$$

To deduce the required result, we show below that T_1 converges in distribution to a $\operatorname{Normal}(0,1)$ random variable and T_2 and T_3 go in probability to 0 and 1, respectively. Slutsky's theorem enables us to combine the results and to obtain the claimed convergence in distribution.

Term T_1: Each degree $d_i = \sum_{j \neq i} A_{ij}$ is a sum of independent random variables. From Assumption 2 ($\Rightarrow \mathbb{E}\, d_i \to \infty$) and Assumption 4 ($\mathbb{E}\, A_{ij} = \Theta(\operatorname{Var} A_{ij})$), it follows that $\operatorname{Var} d_i \to \infty$. Since in addition, the skewness of each edge A_{ij} is asymptotically bounded (Assumption 5), the Lyapunov condition for exponent 1 is satisfied:

$$\frac{\sum_{j \neq i} \mathbb{E}\Big[(A_{ij} - \mathbb{E}\, A_{ij})^3\Big]}{\Big[\sum_{j \neq i} \operatorname{Var} A_{ij}\Big]^{3/2}} \to 0.$$

Hence, the Lindeberg–Feller Central Limit Theorem allows us to conclude that

$$T_1 \xrightarrow{d} \operatorname{Normal}(0,1). \tag{3.6}$$

For more details on the Lindeberg–Feller Central Limit Theorem and the Lyapunov condition see in Appendix A.2 Theorems A.2.5 and A.2.6, respectively.

Term T_2: We write

$$\begin{aligned} T_2 &= \frac{\mathbb{E}\, d_i - \pi_i\sqrt{\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} \\ &= \underbrace{\frac{\mathbb{E}\, d_i - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}}}_{a)} - \underbrace{\frac{\pi_i\sqrt{\|\boldsymbol{d}\|_1} - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}}}_{b)}. \end{aligned} \tag{3.7}$$

Term T_2 converges in probability to 0 since both a) the first ratio converges to 0 and b) the second ratio converges to 0 in probability, as we show below.

$a)$ This convergence is driven by the fact that $\mathbb{E}\, d_i - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1} = \mathcal{O}(1)$ (see Eqs. (3.2) and (3.3)) while $\operatorname{Var} d_i \to \infty$. More precisely, we show in Lemma B.1.1 that under Assumptions 1 and 4 it follows from applying a convergent Taylor expansion that

$$a) \qquad \frac{\mathbb{E}\, d_i - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} = \mathcal{O}\left(\frac{\max_j \pi_j - \pi_i}{\sqrt{n}}\right). \tag{3.8}$$

Since $\pi_j = o(\sqrt{n})$ for all j (Assumption 3), it follows that the left-hand side of Eq. (3.8) converges to 0 in n. $b)$ We show now that the second ratio $\left(\pi_i\sqrt{\|\boldsymbol{d}\|_1} - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}\right)/\sqrt{\operatorname{Var} d_i}$ in Eq. (3.7) converges in probability to 0. First, $\pi_i/\sqrt{\operatorname{Var} d_i} \to 0$ under Assumptions 1 and 4 because

$$\begin{aligned}\frac{\pi_i}{\sqrt{\operatorname{Var} d_i}} &= \Theta\left(\frac{\pi_i}{\sqrt{\mathbb{E}\, d_i}}\right) = \Theta\left(\sqrt{\frac{\pi_i}{\|\boldsymbol{\pi}\|_1 - \pi_i}}\right) \qquad \text{(Assumption 4)} \\ &= \mathcal{O}\big(1/\sqrt{n}\big). \qquad \text{(Assumption 1)}\end{aligned} \tag{3.9}$$

Second, we show in Lemma B.1.2 that the numerator of term b) in Eq. (3.7) is bounded in probability; i.e.,

$$\sqrt{\|\boldsymbol{d}\|_1} - \sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1} = \mathcal{O}_P(1). \tag{3.10}$$

Combining the results in Eqs. (3.8) and (3.10), we conclude that

$$b) \qquad \frac{\pi_i\sqrt{\|\boldsymbol{d}\|_1} - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} \xrightarrow{P} 0.$$

In turn, this completes the proof of the convergence of Term 2 (see Eq. (3.7)); i.e.,

$$T_2 = \underbrace{\frac{\mathbb{E}\, d_i - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}}}_{a)} - \underbrace{\frac{\pi_i\sqrt{\|\boldsymbol{d}\|_1} - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}}}_{b)} \xrightarrow{P} 0. \tag{3.11}$$

Term T_3: From Eqs. (B.4) and (B.5), we know that under Assumptions 2 and 4 it holds that $\|\boldsymbol{d}\|_1/\,\mathbb{E}\,\|\boldsymbol{d}\|_1 = 1 + \mathcal{O}_P\left(\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}/\,\mathbb{E}\,\|\boldsymbol{d}\|_1\right)$ and $\operatorname{Var}\|\boldsymbol{d}\|_1/\,\mathbb{E}\,\|\boldsymbol{d}\|_1 = \Theta(1)$. Thus,

$$\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} = 1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}\right). \tag{3.12}$$

The right-hand-sight of Eq. (3.12) converges in probability to 1 because of Assumption 2 ($\Rightarrow \mathbb{E}\|\boldsymbol{d}\|_1 \to \infty$).

Applying the continuous mapping theorem, it follows that $\sqrt{\|\boldsymbol{d}\|_1 / \mathbb{E}\,\|\boldsymbol{d}\|_1} \xrightarrow{P} 1$. The inverse of a random variable which converges in probability to a constant c, must in turn converge to $1/c$, as long as $c \neq 0$ [85, Theorem 2.1.3]. Thus,

$$T_3 = \sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1}} \xrightarrow{P} 1. \tag{3.13}$$

Slutsky's Theorem (see Appendix A.2) enables us to combine the results on the convergence of terms T_1–T_3 from Eqs. (3.6), (3.11) and (3.13) to obtain that

$$\frac{\hat{\pi}_i - \pi_i}{\sqrt{\operatorname{Var} d_i / \mathbb{E}\,\|\boldsymbol{d}\|_1}} = [T_1 + T_2] \cdot T_3 \to \operatorname{Normal}(0, 1).$$

To complete the proof of Theorem 3.2.1, it remains to show that $\operatorname{Var} d_i / \mathbb{E}\,\|\boldsymbol{d}\|_1 = \mathcal{O}(1/n)$, and that it can be consistently estimated using a plug-in estimator for $A_{ij} \sim \operatorname{Bernoulli}(\pi_i \pi_j)$ and $A_{ij} \sim \operatorname{Poisson}(\pi_i \pi_j)$.

Since $\operatorname{Var} A_{ij} / \mathbb{E}\, A_{ij} = \Theta(1)$ (Assumption 4), we know that

$$\begin{aligned}
\sqrt{\frac{n \operatorname{Var} d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}} &= \Theta\left(\sqrt{\frac{n\ \mathbb{E}\, d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}}\right) \\
&= \Theta\left(\sqrt{\frac{n\ \pi_i(\|\boldsymbol{\pi}\|_1 - \pi_i)}{\|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2}}\right) \\
&= \Theta\left(\sqrt{\frac{n\pi_i}{\|\boldsymbol{\pi}\|_1} \frac{1 - \pi_i/\|\boldsymbol{\pi}\|_1}{1 - \|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2}}\right).
\end{aligned}$$

We know that $n\pi_i/\|\boldsymbol{\pi}\|_1 = \mathcal{O}(1)$ (Assumption 1) and we have seen in Eq. (B.3) that $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$ (from Assumption 1). Hence, we obtain the required result that

$$\sqrt{\frac{\operatorname{Var} d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}} = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

We defer the proof of consistency of the plug-in estimator of $\operatorname{Var} d_i / \mathbb{E}\,\|\boldsymbol{d}\|_1$ for $A_{ij} \sim \operatorname{Bernoulli}(\pi_i \pi_j)$ and $A_{ij} \sim \operatorname{Poisson}(\pi_i \pi_j)$ to Theorem 3.2.2, where we show a more general statement. □

3.2.2 A confidence interval for the estimator of a node's centrality

To quantify the uncertainty of the estimator $\hat{\pi}_i$ of a node's centrality, we show in this section that $\sqrt{\operatorname{Var} d_i / \mathbb{E}\,\|\boldsymbol{d}\|_1}$ can be consistently estimated using a plug-in estimator s by substituting $\hat{\pi}_i$ for π_i in $\mathbb{E}\, d_i$ and $\operatorname{Var} d_i$. In combination with Theorem 3.2.1 it yields a $(1 - \alpha)$-confidence interval (CI) for $\hat{\pi}_i$:

$$\mathrm{CI} = (\hat{\pi}_i - \delta, \hat{\pi}_i + \delta),$$

with $\delta = z_{1-\alpha/2} \cdot s$, and $z_{1-\alpha/2}$ being the $\left(1 - \frac{\alpha}{2}\right)$ quantile of a $\mathrm{Normal}(0, 1)$ distribution.

In Theorem 3.2.1, we claim that the consistency of s holds when $A_{ij} \sim \mathrm{Bernoulli}(\pi_i \pi_j)$ and $A_{ij} \sim \mathrm{Poisson}(\pi_i \pi_j)$. In fact, this is true in more general, as we show in Theorem 3.2.2 below. However, we first need to acknowledge that each edge distribution; e.g. $A_{ij} \sim \mathrm{Bernoulli}(\pi_i \pi_j)$ and $A_{ij} \sim \mathrm{Poisson}(\pi_i \pi_j)$; leads to a different variance $\operatorname{Var} d_i$. All of which are included in Definition 2 since we only assume that $\operatorname{Var} A_{ij} = \Theta(\mathbb{E}\, A_{ij})$ by Assumption 4. We now show that the term $\sqrt{\operatorname{Var} d_i / \mathbb{E}\, \|\boldsymbol{d}\|_1}$ can be consistently estimated by a plug-in estimator s, as long as $\operatorname{Var} d_i$ itself can be consistently estimated by a plug-in estimator. More precisely, we have the following.

Theorem 3.2.2 (Weak consistency of the plug-in estimator for $\sqrt{\operatorname{Var} d_i / \mathbb{E}\, \|\boldsymbol{d}\|_1}$)**.** *Consider Assumptions 1, 2 and 4. Define plug-in estimators* $\widehat{\operatorname{Var} d_i}$ *and* $\widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}$ *by exchanging each* π_i *in* $\operatorname{Var} d_i$ *and* $\mathbb{E}\, \|\boldsymbol{d}\|_1$ *by* $\hat{\pi}_i = d_i / \sqrt{\|\boldsymbol{d}\|_1}$*. In addition, assume that*

$$\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} \xrightarrow{P} 1.$$

Then, the plug-in estimator $\sqrt{\widehat{\operatorname{Var} d_i} / \widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}}$ *consistently estimates* $\sqrt{\operatorname{Var} d_i / \mathbb{E}\, \|\boldsymbol{d}\|_1}$*; i.e.,*

$$\frac{\sqrt{\widehat{\operatorname{Var} d_i} / \widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}}}{\sqrt{\operatorname{Var} d_i / \mathbb{E}\, \|\boldsymbol{d}\|_1}} \xrightarrow{P} 1.$$

Proof. We first write

$$\begin{aligned}
\frac{\sqrt{\widehat{\operatorname{Var} d_i} / \widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}}}{\sqrt{\operatorname{Var} d_i / \mathbb{E}\, \|\boldsymbol{d}\|_1}} &= \sqrt{\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i}} \sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\hat{\boldsymbol{\pi}}\|_1^2 - \|\hat{\boldsymbol{\pi}}\|_2^2}} \\
&= \sqrt{\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i}} \sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1^2 - \|\boldsymbol{d}\|_2^2} \|\boldsymbol{d}\|_1} \\
&= \sqrt{\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i}} \sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1}} \left[1 - \frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2}\right]^{-\frac{1}{2}}.
\end{aligned} \tag{3.14}$$

From term T_3 (Eq. (3.13)) in the proof of Theorem 3.2.1, we know that under Assumption 4 ($\operatorname{Var} A_{ij} = \Theta(\mathbb{E}\, A_{ij})$) it holds that $\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1 / \|\boldsymbol{d}\|_1} \xrightarrow{P} 1$. Since we assume $\widehat{\operatorname{Var} d_i} / \operatorname{Var} d_i \xrightarrow{P} 1$, it remains to show that $\|\boldsymbol{d}\|_2^2 / \|\boldsymbol{d}\|_1^2 \xrightarrow{P} 0$.

We now sketch why $\|\boldsymbol{d}\|_2^2 / \|\boldsymbol{d}\|_1^2 \xrightarrow{P} 0$. For a detailed derivation see Lemma B.1.3 in Appendix B.1.2. Since $\|\boldsymbol{d}\|_1^2$ and $\|\boldsymbol{d}\|_2^2$ concentrate around their respective expectations as $n \to \infty$ (Chebyshev's inequality), and by controlling the error term using a convergent Taylor

expansion, we obtain that

$$\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2} = \frac{\mathbb{E}\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right)\right]. \tag{3.15}$$

Via straightforward algebraic computations, we show that under Assumptions 1, 2, and 4 it holds that

$$\mathbb{E}\|\boldsymbol{d}\|_2^2 = \|\boldsymbol{\pi}\|_1^2\|\boldsymbol{\pi}\|_2^2 \cdot (1 + o(1)), \qquad \mathbb{E}\|\boldsymbol{d}\|_1^2 = \Theta\Big[\big(\mathbb{E}\|\boldsymbol{d}\|_1\big)^2\Big]. \tag{3.16}$$

Combining Eqs. (3.15) and (3.16) and applying Assumption 1, it then follows that

$$\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2} = \mathcal{O}_P\left(\frac{1}{n}\right).$$

Finally, we know from Eq. (3.14) that

$$\frac{\sqrt{\widehat{\operatorname{Var} d_i}/\widehat{\mathbb{E}\,\|\boldsymbol{d}\|_1}}}{\sqrt{\operatorname{Var} d_i/\,\mathbb{E}\,\|\boldsymbol{d}\|_1}} = \sqrt{\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i}}\sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1}}\left[1 - \frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2}\right]^{-\frac{1}{2}}.$$

The inverse of a random variable which converges in probability to a constant c must in turn converge to $1/c$, as long as $c \neq 0$ [85, Theorem 2.1.3]. Applying this fact and the continuous mapping theorem, we obtain the claimed convergence in probability; i.e.,

$$\frac{\sqrt{\widehat{\operatorname{Var} d_i}/\widehat{\mathbb{E}\,\|\boldsymbol{d}\|_1}}}{\sqrt{\operatorname{Var} d_i/\,\mathbb{E}\,\|\boldsymbol{d}\|_1}} \xrightarrow{P} 1.$$

□

Having established Theorem 3.2.2, we may conclude for $A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j)$ and $A_{ij} \sim \text{Poisson}(\pi_i\pi_j)$ that we can consistently estimate $\sqrt{\operatorname{Var} d_i/\,\mathbb{E}\,\|\boldsymbol{d}\|_1}$ using a plug-in estimator since in both cases it holds that $\widehat{\operatorname{Var} d_i}/\operatorname{Var} d_i \xrightarrow{P} 1$, as we show in Lemmas B.1.4 and B.1.5 in Appendix B.1.2.

3.2.3 Multivariate limit theorem

Having shown an univariate central limit theorem for each $\hat{\pi}_i$, we are now ready to extend this result to the multivariate case, applying the Cramér–Wold theorem (see Appendix A.2.3). This result is essential for the analysis of $\widehat{\mathbb{E}\,A_{ij}} = \hat{\pi}_i\hat{\pi}_j$ in Section 3.3.

Corollary 3.2.1 (Multivariate extension for Theorem 3.2.1)**.** *Assume the model of Definition 2 and any finite set of estimators $\hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}$ from Eq.* (3.1)*. Relabeling the indices of these estimators from 1 to r without loss of generality, we have that as $n \to \infty$,*

$$\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}\left(\frac{\hat{\pi}_1 - \pi_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\hat{\pi}_r - \pi_r}{\sqrt{\operatorname{Var} d_r}}\right) \xrightarrow{d} \text{Normal}(0, \boldsymbol{I}_r).$$

Proof. This proof is the multidimensional equivalent of the proof of Theorem 3.2.1. It is analogously driven by the fact that the vector

$$\boldsymbol{m_1} = \left(\frac{d_1 - \mathbb{E}\, d_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{d_r - \mathbb{E}\, d_r}{\sqrt{\operatorname{Var} d_r}} \right)'$$

can be reduced to a sum of independent but not identically distributed random vectors. These in turn converge in distribution to a multivariate standard Normal random vector; as we now show. In direct analogy to the univariate case of Eq. (3.5),

$$\begin{aligned}
&\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1} \left(\frac{\hat{\pi}_1 - \pi_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\hat{\pi}_r - \pi_r}{\sqrt{\operatorname{Var} d_r}} \right)' \\
&= \sqrt{\mathbb{E}\|\boldsymbol{d}\|_1} \left(\frac{1}{\sqrt{\operatorname{Var} d_1}} \left(\frac{d_1}{\sqrt{\|\boldsymbol{d}\|_1}} - \pi_1 \right), \ldots, \frac{1}{\sqrt{\operatorname{Var} d_r}} \left(\frac{d_r}{\sqrt{\|\boldsymbol{d}\|_1}} - \pi_r \right) \right)' \\
&= \underbrace{\sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\|\boldsymbol{d}\|_1}}}_{m_3} \cdot \Bigg(\underbrace{\left(\frac{d_1 - \mathbb{E}\, d_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{d_r - \mathbb{E}\, d_r}{\sqrt{\operatorname{Var} d_r}} \right)'}_{\boldsymbol{m_1}} \\
&\qquad + \underbrace{\left(\frac{\mathbb{E}\, d_1 - \pi_1 \sqrt{\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\mathbb{E}\, d_r - \pi_r \sqrt{\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_r}} \right)'}_{\boldsymbol{m_2}} \Bigg).
\end{aligned} \tag{3.17}$$

Each component of the vector $\boldsymbol{m_2}$ converges in probability to $\mathbf{0}$ (see Eq. (3.11) in the proof of Theorem 3.2.1). It follows that the vector $\boldsymbol{m_2} \xrightarrow{P} \mathbf{0}$. In addition, the scalar m_3 converges in probability to 1 (see Eq. (3.13) in the proof of Theorem 3.2.1).

We now prove that $\boldsymbol{m_1} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r})$. In order to apply a multivariate central limit theorem, we rearrange $\boldsymbol{m_1}$ such that we extract a sum of independent random vectors ($\boldsymbol{m_{12}}$):

$$\begin{aligned}
\boldsymbol{m_1} &= \left(\frac{d_1 - \mathbb{E}\, d_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{d_r - \mathbb{E}\, d_r}{\sqrt{\operatorname{Var} d_r}} \right)' \\
&= \underbrace{\operatorname{diag}\left(\frac{\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{l1})}}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{lr})}}{\sqrt{\operatorname{Var} d_r}} \right)}_{\boldsymbol{D_{11}}} \\
&\quad \cdot \underbrace{\left(\frac{\sum_{l=r+1}^{n} (A_{l1} - \mathbb{E}\, A_{l1})}{\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{l1})}}, \ldots, \frac{\sum_{l=r+1}^{n} (A_{lr} - \mathbb{E}\, A_{lr})}{\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{lr})}} \right)'}_{\boldsymbol{m_{12}}} \\
&\quad + \underbrace{\left(\frac{\sum_{l=1}^{r} (A_{l1} - \mathbb{E}\, A_{l1})}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\sum_{l=1}^{r} (A_{lr} - \mathbb{E}\, A_{lr})}{\sqrt{\operatorname{Var} d_r}} \right)'}_{\boldsymbol{m_{13}}}.
\end{aligned} \tag{3.18}$$

We show in Lemma B.1.6 in Appendix B.1.3 that given Assumptions 1–5, the following three things hold. First, the matrix $\boldsymbol{D_{11}}$ converges to the identity matrix $\boldsymbol{I_r}$ because the finite number

of summands that the numerator is short compared to the denominator is negligible. Second, using the Lindeberg–Feller central limit theorem with the Lyapunov condition we show that each component of $\boldsymbol{m_{12}}$ converges marginally to a $\mathrm{Normal}(0, 1)$ random variable; since the components are independent we can apply the Cramér–Wold theorem (see Appendix A.2) to conclude $\boldsymbol{m_{12}} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r})$. Third, from Chebyshev's inequality it follows that the term $\boldsymbol{m_{13}} \xrightarrow{P} \mathbf{0}$.

By Slutsky's theorem (see Appendix A.2), we can combine the results on the convergence of $\boldsymbol{D_{11}}$, $\boldsymbol{m_{12}}$, and $\boldsymbol{m_{13}}$ to conclude (see Eq. (3.18)) that

$$\boldsymbol{m_1} = \boldsymbol{D_{11}}\, \boldsymbol{m_{12}} + \boldsymbol{m_{13}} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r}).$$

In turn, we deduce the overall required result of this corollary (see Eq. (3.17)) that

$$\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}\left(\frac{\hat{\pi}_1 - \pi_1}{\sqrt{\operatorname{Var} d_1}}, \ldots, \frac{\hat{\pi}_r - \pi_r}{\sqrt{\operatorname{Var} d_r}}\right)' = m_3 \boldsymbol{m_1} + \boldsymbol{m_2} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r}).$$

□

Recall that for all i, $\sqrt{\operatorname{Var} d_i / \mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}(1/\sqrt{n})$, and we can consistently estimate it if $A_{ij} \sim \mathrm{Bernoulli}(\pi_i \pi_j)$ or $A_{ij} \sim \mathrm{Poisson}(\pi_i \pi_j)$ by substituting $\hat{\pi}_i$ for π_i in $\mathbb{E}\, d_i$ and $\operatorname{Var} d_i$ for all i. Thus, in combination with Corollary 3.2.1 we now can quantify the uncertainty to estimate $(\pi_1, \ldots, \pi_r)$ as $(\hat{\pi}_1, \ldots, \hat{\pi}_r)$ by a multivariate confidence interval.

3.3 Properties of the estimator of an edge expectation

We may apply the results of Section 3.2 to characterize an estimator for the edge expectations $\mathbb{E}\, A_{ij}$, for all i, j. From Definition 2 ($\mathbb{E}\, A_{ij} = \pi_i \pi_j$, beside others) and Eq. (3.1) $\left(i.e., \hat{\pi}_i = d_i / \sqrt{\|\boldsymbol{d}\|_1}\right)$, it is natural to define an estimator for $\mathbb{E}\, A_{ij}$ as

$$\widehat{\mathbb{E}\, A}_{ij} = \hat{\pi}_i \hat{\pi}_j = \frac{d_i d_j}{\|\boldsymbol{d}\|_1}, \quad 1 \le i < j \le n. \tag{3.19}$$

While deriving modularity from first principles in Chapter 4, we will understand that it estimates the expectation of an edge as in Eq. (3.19). Hence, to understand the strengths and weaknesses of modularity, we need to derive the properties of the estimators $\widehat{\mathbb{E}\, A}_{ij}$.

3.3.1 Weak consistency of the estimator of an edge expectation

First of all, the estimator in Eq. (3.19) is weakly consistent for the edge expectation $\mathbb{E}\, A_{ij}$ under Definition 2. To be more precise, we obtain the following.

Corollary 3.3.1 (Weak consistency for Eq. (3.19))**.** *Consider Assumptions 1, 2, and 4. Then, for any $i, j \in \mathbb{N}_{>0}$ it holds that*

$$\frac{\widehat{\mathbb{E}\, A_{ij}}}{\mathbb{E}\, A_{ij}} \xrightarrow{P} 1.$$

Proof. It can easily be seen that

$$\widehat{\mathbb{E}\, A_{ij}} - \mathbb{E}\, A_{ij} = \pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j) + (\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j). \tag{3.20}$$

Furthermore, we know from Lemma B.1.7 in Appendix B.1.4 that under Assumptions 1, 2, and 4 it holds that

$$= (\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)) \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}}\right)\right].$$

From Lemma B.1.7 and Assumptions 1, 2 and 4, it follows that

$$\begin{aligned}
&= \mathcal{O}_P\left(\frac{\pi_i\pi_j}{\sqrt{\mathbb{E}\, d_i}} + \frac{\pi_i\pi_j}{\sqrt{\mathbb{E}\, d_j}}\right) \\
&= \mathcal{O}_P\left(\sqrt{\frac{\pi_i}{\|\boldsymbol{\pi}\|_1}}\pi_j + \sqrt{\frac{\pi_j}{\|\boldsymbol{\pi}\|_1}}\pi_i\right) \quad \text{(Assumption 1)} \\
&= \mathcal{O}_P\left(\frac{\pi_j}{\sqrt{n}} + \frac{\pi_i}{\sqrt{n}}\right). \quad \text{(Assumption 1)}
\end{aligned}$$

Thus, we can conclude the required result since

$$\begin{aligned}
\frac{\widehat{\mathbb{E}\, A_{ij}}}{\mathbb{E}\, A_{ij}} - 1 &= \mathcal{O}_P\left(\frac{\pi_j}{\pi_i\pi_j\sqrt{n}} + \frac{\pi_i}{\pi_i\pi_j\sqrt{n}}\right) \\
&= \mathcal{O}_P\left(\frac{1}{\pi_i\sqrt{n}} + \frac{1}{\pi_j\sqrt{n}}\right) \\
&= o_P(1). \quad \text{(Assumption 2)}
\end{aligned}$$

□

3.3.2 A limit theorem for the estimator of an edge expectation

In the previous section, we show that we can consistently estimate $\mathbb{E}\, A_{ij}$ using $\widehat{\mathbb{E}\, A_{ij}}$ if the network is generated from a degree-based model of Definition 2 (Corollary 3.2.1). In this section, as a consequence, we derive the large sample distribution of $\widehat{\mathbb{E}\, A_{ij}}$ to fully describe its asymptotic behavior under a degree-based model.

Corollary 3.3.2 (Central limit theorem for Eq. (3.19))**.** *Assume the model of Definition 2 and define an estimator* $\widehat{\mathbb{E}\,A}_{ij}$ *of* $\mathbb{E}\,A_{ij}$ *as in Eq.* (3.19)*. Then as* $n \to \infty$*,*

$$\frac{\widehat{\mathbb{E}\,A}_{ij} - \mathbb{E}\,A_{ij}}{\sqrt{\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \mathbb{E}\|\boldsymbol{d}\|_1}} \xrightarrow{d} \operatorname{Normal}(0,1),$$

for any $i, j \in \mathbb{N}_{>0}$*. Furthermore,* $\sqrt{\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}\big(\sqrt{\mathbb{E}\,A_{ij}/n}\big)$*, and can be consistently estimated if* $A_{ij} \sim \operatorname{Bernoulli}(\pi_i\pi_j)$ *or* $A_{ij} \sim \operatorname{Poisson}(\pi_i\pi_j)$ *by substituting* $\hat{\boldsymbol{\pi}}$ *for* $\boldsymbol{\pi}$*; for any* i, j*.*

Proof. We show that $\widehat{\mathbb{E}\,A}_{ij} = \hat{\pi}_i\hat{\pi}_j$, once appropriately standardized, converges in distribution to a $\operatorname{Normal}(0,1)$ random variable. Recall from Eq. (3.20) that

$$\widehat{\mathbb{E}\,A}_{ij} = \pi_i\pi_j + \pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j) + (\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j). \tag{3.21}$$

Since $(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)$ is asymptotically negligible as we show in Lemma B.1.7 in Appendix B.1.4, the asymptotic behavior of $\widehat{\mathbb{E}\,A}_{ij} - \pi_i\pi_j$ is dominated by $\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)$. As a consequence, we standardize all quantities of both sides of Eq. (3.21) by the factor $\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1 / \left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right)}$, which can be interpreted as an approximation of the standard deviation of $\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)$. Then, we can use Eq. (3.21) to write

$$\begin{aligned}
&\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}\,\frac{\widehat{\mathbb{E}\,A}_{ij} - \pi_i\pi_j}{\sqrt{\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j}} \\
&= \underbrace{\sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j}} \left[\pi_j \sqrt{\operatorname{Var} d_i}\left(\frac{\hat{\pi}_i - \pi_i}{\sqrt{\operatorname{Var} d_i}}\right) + \pi_i \sqrt{\operatorname{Var} d_j}\left(\frac{\hat{\pi}_j - \pi_j}{\sqrt{\operatorname{Var} d_j}}\right)\right]}_{T_1} \\
&\quad + \underbrace{\sqrt{\frac{\mathbb{E}\|\boldsymbol{d}\|_1}{\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j}} \cdot (\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}_{T_2}.
\end{aligned} \tag{3.22}$$

To deduce the required result, we show that $T_1 \xrightarrow{d} \operatorname{Normal}(0,1)$ and that $T_2 = o_P(T_1)$. Slutsky's theorem will then enable us to combine these results and obtain the claimed convergence in distribution.

Term T_1: Recall from Corollary 3.2.1 that under Assumptions 1–5 it holds that $\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}\left(\frac{\hat{\pi}_i - \pi_i}{\sqrt{\operatorname{Var} d_i}}, \frac{\hat{\pi}_j - \pi_j}{\sqrt{\operatorname{Var} d_j}}\right)' \xrightarrow{d} \operatorname{Normal}(\mathbf{0}, \boldsymbol{I_2})$. Applying the Cramér–Wold theorem (see Appendix A.2), we can conclude that

$$T_1 \xrightarrow{d} \operatorname{Normal}(0,1). \tag{3.23}$$

Term T_2: Recall from Lemma B.1.7 in Appendix B.1.4 that

$$\frac{T_2}{T_1} = \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}}\right).$$

From Assumption 2 ($\pi_i = \omega(1/\sqrt{n}), \forall i$), it follows that $\mathbb{E}\, d_i$ diverges $\forall i$, and hence that

$$\frac{T_2}{T_1} \xrightarrow{P} 0. \tag{3.24}$$

Combining the results in Eqs. (3.22), (3.23), and (3.24), we obtain the required result:

$$\frac{\widehat{\mathbb{E}\, A_{ij}} - \mathbb{E}\, A_{ij}}{\sqrt{\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \mathbb{E}\|\boldsymbol{d}\|_1}} = T_1 + T_2 \xrightarrow{d} \text{Normal}(0,1).$$

To complete the proof, two more steps remain to be shown. First, it holds that $\sqrt{\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}\big(\sqrt{\mathbb{E}\, A_{ij}/n}\big) = \mathcal{O}\big(\sqrt{\pi_i \pi_j / n}\big)$:

$$\begin{aligned}
&\sqrt{\frac{n}{\pi_i \pi_j} \cdot \frac{\pi_i^2 \operatorname{Var} d_j + \pi_j^2 \operatorname{Var} d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}} \\
&= \Theta\left(\sqrt{\frac{n}{\pi_i \pi_j} \cdot \frac{\pi_i^2 \,\mathbb{E}\, d_j + \pi_j^2 \,\mathbb{E}\, d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}}\right) \quad \text{(Assumption 4)} \\
&= \Theta\left(\sqrt{\frac{n}{\pi_i \pi_j} \cdot \frac{\pi_i^2 \pi_j + \pi_j^2 \pi_i}{\|\boldsymbol{\pi}\|_1}}\right) \quad \text{(Assumption 1)} \\
&= \Theta\left(\sqrt{n \cdot \frac{\pi_i + \pi_j}{\|\boldsymbol{\pi}\|_1}}\right) \\
&= \mathcal{O}(1). \quad \text{(Assumption 1)}
\end{aligned}$$

Second, we show in Appendix A.2 that the consistency of the plug-in estimator of $\sqrt{n\left(\pi_j^2 \operatorname{Var} d_i + \pi_i^2 \operatorname{Var} d_j\right) / \mathbb{E}\|\boldsymbol{d}\|_1}$ for networks with edges $A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j)$ or $A_{ij} \sim \text{Poisson}(\pi_i \pi_j)$ follows from Lemma B.1.7 in Appendix B.1.4 $\big($i.e., $\hat{\pi}_i/\pi_i - 1 = \mathcal{O}_P\big(1/\sqrt{\mathbb{E}\, d_i}\big)\big)$ and Theorem 3.2.2 (i.e., the consistency of the plug-in estimator of $\sqrt{\operatorname{Var} d_i / \mathbb{E}\|\boldsymbol{d}\|_1}$), using Slutsky's Theorem. □

3.4 Illustrative simulations

We run several simulations to illustrate the theoretical results of Theorems 3.2.1 and 3.2.2, and Corollary 3.3.2 on the asymptotic properties of the estimator of a node's centrality $\hat{\pi}_i$ and of the estimator of an edge expectation $\widehat{\mathbb{E}\, A_{ij}}$. Since the simulations discussing Corollary 3.3.2

are very similar in set-up, outcome and interpretation as the simulations for Theorem 3.2.1, we postpone them to Appendix B.2.3 for the more curious readers. For this section, we focus on simple random graphs that follow Definition 3.1, where we model the edges as Bernoulli random variables with success probability $\pi_i\pi_j$:

$$A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j).$$

For simulations that illustrate the variety of models included in Definition 3.1 see the illustrative simulations in Chapter 4.

3.4.1 Illustrative simulations for the limit theorem for the estimator of a node's centrality

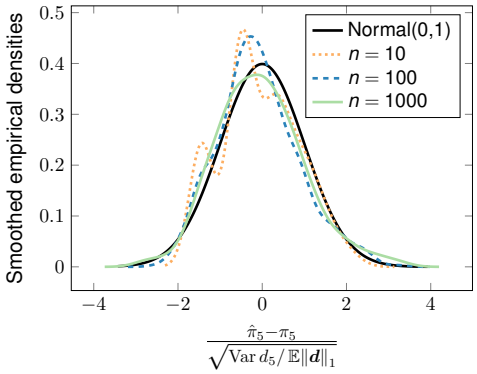
In many observed networks, the sorted degrees behave similarly to a power law; e.g., the internet, social, and citation networks (see Section 1.3.1 or [48, p. 11]). We therefore decide to generate the parameters π_i as $\pi_i = \theta\ i^{-\gamma}$, for $i = 1, \ldots, n$ which leads to networks where the expected degrees follow a power law. As mentioned above, in applications scientists observe that the proportion of nodes with $d_i = k$ scales approximately as $k^{-\beta}$ and they report the empirical exponent to lie typically between $2 < \beta < 3$ [10, 32]. These two models (i.e., $\theta\ i^{-\gamma}$ and $k^{-\beta}$) have been related in [116] as $\gamma = 1/(\beta - 1)$ and in turn we simulate from a degree-based model with Bernoulli($\pi_i\pi_j$) edges and parameters:

$$\pi_i = \theta\ i^{-\gamma} \qquad \text{for } 1/2 < \gamma < 1.$$

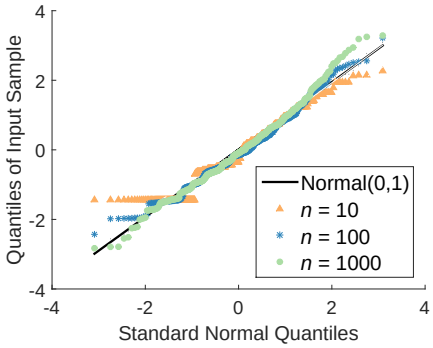
We draw 500 independent realizations from this model and estimate for each of these networks the model parameter of a specific node i, π_i. We standardize the estimator $\hat{\pi}_i$ by its approximated expectation π_i and standard deviation, leading to

$$\frac{\hat{\pi}_i - \pi_i}{\sqrt{\text{Var}\, d_i / \mathbb{E}\|\boldsymbol{d}\|_1}}.$$

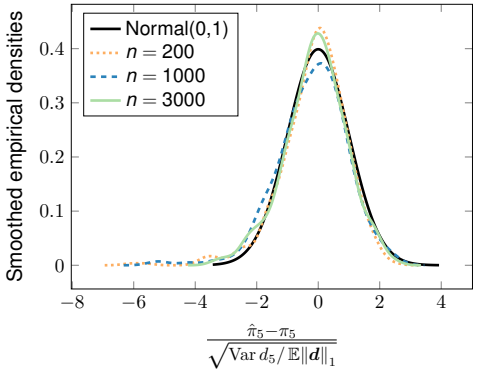
As a result, we illustrate in Figure 3.1 that in agreement with Theorem 3.2.1 an increase in number of nodes n leads to an improvement of the $\text{Normal}(0, 1)$ approximation for the empirical density of the estimator $\hat{\pi}_i$. Figure 3.1a displays the smoothed empirical density of $\hat{\pi}_5$ for 500 repetitions from a degree-based model with $\pi_i = i^{-0.6}$ for all i. For comparison, we added a $\text{Normal}(0, 1)$ density. We compute the smoothed empirical density using a Gaussian kernel estimator. In Figure 3.1b, we display the quantile-quantile plots (Q-Q plot) of the standardized estimator to illustrate that the empirical densities do not only appear to be $\text{Normal}(0, 1)$ due to the Gaussian kernel estimator.



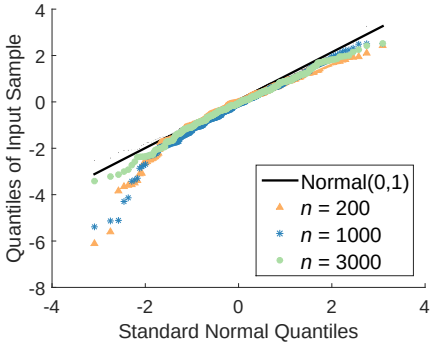
(a) Smoothed empirical densities



(b) Q-Q plot



(c) Plug-in: Smoothed empirical densities



(d) Plug-in: Q-Q plot

Figure 3.1: Illustration of Theorem 3.2.1: The large-sample behavior of the estimator of the centrality of node 5; simulated from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.6}$.

In practice, we cannot observe π_j for any $j = 1, \ldots, n$. Thus, we estimate the approximated standard deviation of $\hat{\pi}_i$ using a plug-in estimator of $\sqrt{\mathrm{Var}\, d_i / \mathbb{E}\|\boldsymbol{d}\|_1}$ where we substitute all π_js by $\hat{\pi}_j$s. Figures 3.1c and 3.1d show the same plots for $\hat{\pi}_5$ as Figures 3.1a and 3.1b but standardized by the plug-in estimator for $\sqrt{\mathrm{Var}\, d_i / \mathbb{E}\|\boldsymbol{d}\|_1}$. We observe that using the plug-in estimator instead of the true value of $\sqrt{\mathrm{Var}\, d_i / \mathbb{E}\|\boldsymbol{d}\|_1}$, slows down the convergence in distribution in n.

The convergence in distribution is not driven by the number of nodes but the effective sample size. We repeat the simulations described above for varied $\theta \in (0, 1]$, $\gamma = [0, 1)$ and node indices i. In all cases we observe that as n increases the difference between the smoothed empirical density of π_i and a $\mathrm{Normal}(0, 1)$ density shrinks. However, we see that if the network is less sparse; e.g. $\gamma < 0.6$ it takes fewer nodes to see the same convergence rate. Figures B.1 and B.2 in Appendix B.2.1 illustrate this point with simulations from $\pi_i = i^{-0.2}$ and $\pi_i = 0.9\, i^{-0.2}$ for node 5 and 17, respectively.

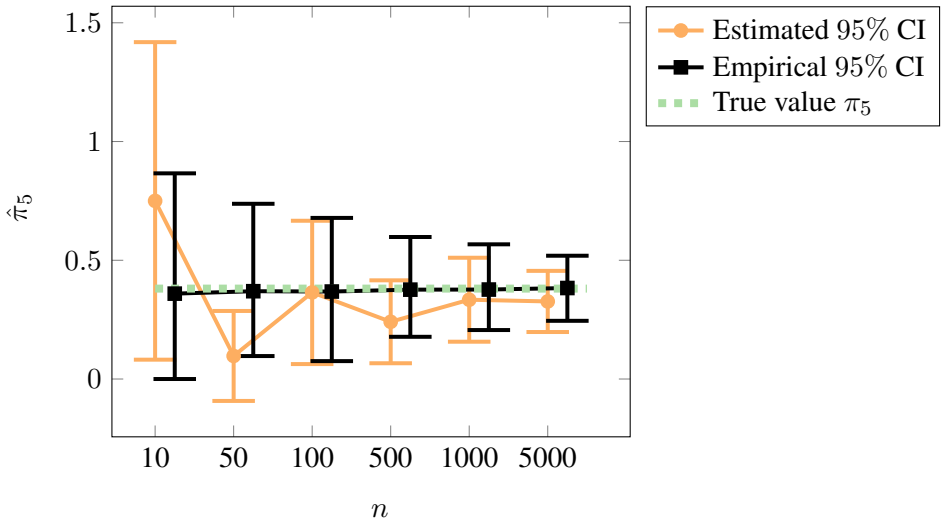


Figure 3.2: The estimator $\hat{\pi}_5$, shown along with its estimated and empirical large-sample confidence intervals; from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.6}$.

3.4.2 Illustrative simulations for the confidence interval for the estimator of a node's centrality

To illustrate the quality of the estimated confidence interval that results from combining Theorems 3.2.1 and 3.2.2, we compare the estimated confidence interval with an empirical confidence interval based on simulations, for several numbers of nodes n. Based on a single network, we estimate the confidence interval as

$$\text{Estimated CI} = (\hat{\pi}_i - \delta, \hat{\pi}_i + \delta), \quad \delta = z_{1-\alpha/2} \cdot \sqrt{\widehat{\operatorname{Var}}\, d_i / \widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}},$$

where $z_{1-\alpha/2}$ denotes the $\left(1 - \frac{\alpha}{2}\right)$ quantile of a $\text{Normal}(0, 1)$ distribution and $\widehat{\operatorname{Var}}\, d_i$ and $\widehat{\mathbb{E}\, \|\boldsymbol{d}\|_1}$ the plug-in estimators of $\operatorname{Var} d_i$ and $\mathbb{E}\, \|\boldsymbol{d}\|_1$ (as before). In contrast, we derive the empirical confidence interval based on 1000 simulations from the same power law network model. To be more precise, we compute the 2.5% and 97.5% quantiles $q_{2.5\%}$ and $q_{97.5\%}$ centering the interval at the empirical mean such that

$$\text{Empirical CI} = (q_{2.5\%}, q_{97.5\%}).$$

In Figure 3.2, we display the results for node 5 from simulations from a power law network with $\pi_i = i^{-0.6}$. In addition to the estimated and empirical confidence intervals, we display the true value $\pi_5 = 0.3807$ as the target. In agreement with Theorem 3.2.1, we observe that as n increases the empirical confidence interval shrinks; indicating that the estimators $\hat{\pi}_5$ improve as

n goes to infinity. We also see that the center of the confidence interval—the empirical mean—matches the true value even for small networks. As the empirical mean is a common estimator for the expectation, it indicates that $\hat{\pi}_i$ may be asymptotically unbiased with a fast convergence rate. However, the width of the empirical confidence intervals indicates that for small networks ($n \leq 500$) the variance of the estimators $\hat{\pi}_i$ may be large.

In agreement with Theorem 3.2.2, as the number of nodes n increases the overlap between the estimated confidence interval with the empirical confidence interval increases. The empirical confidence interval is centered at the empirical mean and its length is scaled with the empirical standard deviation of $\hat{\pi}_i$—good estimators for the true mean and standard deviation of $\hat{\pi}_i$. Here, good means unbiased and consistent with fast convergence. Thus, the empirical confidence intervals give a reliable description of the estimator $\hat{\pi}_i$. However, to compute the empirical confidence interval we need about 1000 independent realizations of the network. Thus, the observation that the estimated confidence interval (based on a single realization) increasingly overlaps with the empirical confidence interval for growing n; indicates that the estimated confidence interval is of good and, as $n \to \infty$, improving quality.

In addition, the estimated 95% confidence interval in Figure 3.2 shrinks as the number of nodes n increases; illustrating that as larger the network as more informative is the estimated confidence interval. Furthermore, we observe that for small networks ($n \leq 500$) the estimated confidence interval does not necessarily include the true value. However, as n increases the confidence interval becomes more and more reliable in covering the true value.

We repeat the simulation study for several different values of θ, γ and the node index i. In all cases, we observe that the estimated confidence interval becomes more informative and more reliable as the number of nodes n increases. Similar to the previous simulations in Section 3.4.1, we observe that as the network becomes more dense the estimated confidence interval shrinks faster and overlaps earlier (in n) with the empirical confidence interval. See a few examples for illustration in Appendix B.2.2.

3.5 Discussion

In this chapter, we introduced a new family of network models that naturally extends the degree-based model of Chung and Lu [30] to include beside others weighted and multi-edge networks. This family of models describes the structure of a network solely by the collection of the nodes' centralities (i.e., degree-centralities).

We here derived a central limit theorem for the estimators of the model parameters $\hat{\pi}_i$ in this general setting; thereby extending work in [116]. We know from [119] that these estimators are near-maximum likelihood estimators for sparse, and simple networks. As a consequence of our theoretical results, we are now able to decide whether two nodes are equally central while controlling the Type I error of falsely rejecting the hypothesis of equality. Furthermore, we delivered an estimator for the approximated standard deviations of $\hat{\pi}_i$; allowing us to quantify the uncertainty of our estimators $\hat{\pi}_i$ by a confidence interval.

To discuss modularity in the following chapter, we need to estimate the expectation of an edge $\mathbb{E}\, A_{ij}$ instead of π_i itself, under a degree-based model. We therefore derived a central limit theorem for the estimator of $\mathbb{E}\, A_{ij}$ that naturally follows from $\hat{\pi}_i$ under a degree-based model. As a result, we know that when a degree-based model is in place, the estimators $\widehat{\mathbb{E}\, A_{ij}} = d_i d_j / \|\boldsymbol{d}\|_1$ perform well for large networks.

Chapter 4

Significance of a community structure under degree-based models

We here derive theoretical results enabling us to assess whether an observed community structure is informative for the interactions in a network, while controlling the Type I error of falsely identifying a community assignment as such. As before, we call a community structure informative if there are significantly more edges within than between communities. We do so by way of modularity: a well-known quality measure for community structure that is mainly used for community detection (see Section 2.5). As such, modularity has proven to be useful in practice and over time became a popular community detection method. However, it lacks a theoretical foundation and is not objective since a value may have a different meaning depending on the sparsity and size of the network. In the work presented here, we establish a statistical interpretation of modularity, enabling us to derive its asymptotic distribution under the family of degree-based models. As a consequence, we gain a theoretical understanding of modularity, and may in turn utilize it to translate a network and its observed community structure into a p-value that objectively assesses whether the observed community structure is informative for the interactions in a network.

While most work on community structure in networks focuses on identifying a single "best" community assignment (e.g. latent space models [67], stochastic block models [68], degree-corrected stochastic block models [75, 110], and modularity [106, 153]), we here take a different approach. A network is a simple way to describe the complex structure of interactions between units. We believe that such a structure may not be well described with a single community assignment. Instead, there are different motivations for units to interact best described by multiple community assignments. Furthermore, since networks often come with covariates on the nodes, we infer the community structures from the covariates adding interpretability. In

fact, many datasets nowadays come with several covariates on the nodes [18, 103, 118, 130]; raising the question of which of these covariates reflect a network's structure. We answer this question by delivering a methodology that assesses a covariate-based community structure with a p-value; enabling us to decide whether the community structure is informative while controlling the Type I error.

To derive this methodology, it is crucial to understand the family of the degree-based models; as derived in the previous chapter. With a single parameter per node, these models allow for node-specific differences while lacking the flexibility to support community structure. Therefore, members of this family are frequently used as null models for no community structure. The formulation in Definition 3.1 reduces the properties to those essential for modeling networks with a lack of community structure in the context of modularity. The results in Theorems 3.2.1 and 3.2.2, and Corollary 3.3.2 derived in the previous chapter ensure that when assuming a degree-based model, the estimators for the model parameters $\hat{\pi}_i$, and the edge expectations $\widehat{\mathbb{E}\, A_{ij}}$ are well behaved for large networks.

4.1 Modularity in the presence of observed community structure

Three essential ingredients are necessary to understand modularity in the presence of covariates: first, a formal interpretation of modularity as a measure of statistical significance; second, the use of this framework to evaluate a covariate-based community assignment; and third, the model that is underlying modularity. We now describe each of these ingredients in turn.

First, to interpret modularity as a measure of statistical significance, we must recognize it as an estimator of a population quantity. Recall that $\delta_{g(i)=g(j)} = 1$ when nodes i and j are assigned to the same community, and 0 otherwise. Then, we know from Definition 1 that when $\|\boldsymbol{d}\|_1 \neq 0$, modularity is defined as

$$\widehat{Q} = \sum_{j=1}^{n} \sum_{i<j} \left[A_{ij} - \frac{d_i d_j}{\|\boldsymbol{d}\|_1} \right] \delta_{g(i)=g(j)}. \tag{4.1}$$

Modularity contrasts an observed edge A_{ij} with the ratio $d_i d_j / \|\boldsymbol{d}\|_1$ whenever nodes i and j are in the same community. Now consider replacing $d_i d_j / \|\boldsymbol{d}\|_1$ by $\mathbb{E}\, A_{ij}$, the expected value of an edge under a given model:

$$Q = \sum_{j=1}^{n} \sum_{i<j} [A_{ij} - \mathbb{E}\, A_{ij}] \delta_{g(i)=g(j)}. \tag{4.2}$$

Substituting $\widehat{\mathbb{E}\, A_{ij}} = d_i d_j / \|\boldsymbol{d}\|_1$ (see Eq. (3.19)) for $\mathbb{E}\, A_{ij}$ in Eq. (4.2), we immediately recognize modularity $\widehat{Q}$ as defined in Eq. (4.1). Thus, modularity implicitly assumes a degree-based model of Definition 2.

We recognize Q in Eq. (4.2) as a sum of signed residuals (observed minus expected values) $A_{ij} - \mathbb{E}\, A_{ij}$. If the model for each $\mathbb{E}\, A_{ij}$ posits the *absence* of community structure, then a large positive value of Q indicates the *presence* of such structure (more within-group edges than expected). In Figure 1.7, we have seen this effect already in the student friendship network: the visible community structure in Figures 1.7a–c is obscured in Figure 1.7d when communities are assigned at random. Moreover, using $d_i d_j / \|\boldsymbol{d}\|_1$ as a proxy for $\mathbb{E}\, A_{ij}$, we see that modularity $\widehat{Q}$ as defined in Eq. (4.1) is an estimator of Q in Eq. (4.2). We will return to this point in the next section.

Second, to interpret covariate-based community structure, we must recognize that different community assignments reveal different structural aspects of a network. Figures 1.7a–c illustrate this point by grouping a student friendship network by gender, race, and year in school. Covariates such as these define distinct community assignments, each of which relates the covariate in question to the observed network structure.

A key insight is that rather than maximizing modularity to obtain a single "best" community assignment, we may instead use modularity to measure the strength of an observed community structure. If a particular community assignment is given by a covariate, then modularity allows us to quantify the explanatory value of this covariate for the observed structure of the network.

Third, modularity indirectly assumes a model since different models for the network edges A_{ij} will imply different estimators for Q in Eq. (4.2). Estimating Q using $\widehat{Q}$ in Eq. (4.1), we indirectly assume a model for the absence of community structure, where nodes connect independently based on the product of their individual propensities to form connections. In particular, by using the estimator $\widehat{\mathbb{E}\, A_{ij}} = d_i d_j / \|\boldsymbol{d}\|_1$ as a proxy for $\mathbb{E}\, A_{ij}$, we see that in fact the model underlying modularity is part of the family of degree-based models of Definition 2 for which $\widehat{\mathbb{E}\, A_{ij}}$ is the natural estimator.

Up until now, we derived modularity $\widehat{Q}$ (see Eq. (4.1)) from first principles and gave it a formal statistical interpretation as an estimator of a population quantity Q (see Eq. (4.2)). Furthermore, we have identified the degree-based model of Definition 2 as the model underlying modularity $\widehat{Q}$. From Chapter 3, we know that the estimators of its parameters $\widehat{\mathbb{E}\, A_{ij}}$ behave well for large networks. Building up on these results, we now analyze the asymptotic properties of

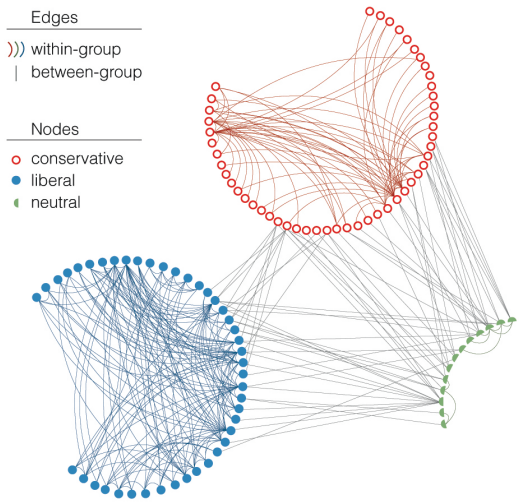


Figure 4.1: Decomposition of a network in within- and between-group edges: political books connected when frequently purchased together, where groups are defined by political alignment [108]. Note that only within-group edges appear in Q (Eq. (4.2)); by contrast, both types of edges contribute to $\widehat{Q}$ (Eq. (4.1)).

modularity $\widehat{Q}$ under a degree-based model.

4.2 Properties of modularity

In this section, we derive the limiting properties of modularity itself, putting it on a sound theoretical basis for the first time. We first decompose modularity into a bias- and a variance-related component (see Section 4.2.1), to then derive its asymptotic distribution (see Section 4.2.2). As a consequence, we can extend its usage to objectively quantify observed network structure.

4.2.1 Modularity reflects within- and between-group edges

Corollary 3.3.2 in Section 3.3 on the asymptotic normality of $\widehat{\mathbb{E}\, A_{ij}}$ leads to the first of two key insights to the asymptotic behavior of modularity: Recall that $\widehat{Q}$ (Eq. (4.1)) is an estimator for its population counterpart Q (Eq. (4.2)), in which $\widehat{\mathbb{E}\, A_{ij}}$ estimates $\mathbb{E}\, A_{ij}$. Comparing Eqs. (4.1) and (4.2), and approximating $\mathbb{E}\big(d_i d_j / \|\boldsymbol{d}\|_1\big)$ by $\mathbb{E}\, d_i d_j / \mathbb{E}\|\boldsymbol{d}\|_1$, we obtain:

$$\mathbb{E}(\widehat{Q} - Q) \approx \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \right) \delta_{g(i)=g(j)}. \tag{4.3}$$

The approximation of $\mathbb{E}\big(d_i d_j / \|\boldsymbol{d}\|_1\big)$ by $\mathbb{E}\, d_i d_j / \mathbb{E}\|\boldsymbol{d}\|_1$ can be quantified using a first order Taylor expansion; leading to convergence in probability. However, this does not imply convergence in moments, and thus Eq. (4.3) is only an approximation. Under the model of Definition 2,

the difference in Eq. (4.3) cancels to first order (see Appendix C.1.1), yielding an approximate bias term of modularity:

$$b = \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} \left(\mathbb{E}\, d_i + \mathbb{E}\, d_j - \|\boldsymbol{\pi}\|_2^2 \right)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)}. \tag{4.4}$$

Figure 4.1 illustrates the second key insight into the limiting behavior of modularity: its variability reduces asymptotically to that of a centered sum of within- and between-group edges. More specifically, every network degree $d_j = \sum_{i \neq j} A_{ij}$ decomposes into within- and between-group components:

$$\begin{gathered} d_j = d_j^w + d_j^b; \\ d_j^w = \sum_{i \neq j} A_{ij} \delta_{g(i)=g(j)}, \qquad d_j^b = \sum_{i=1}^{n} A_{ij} \delta_{g(i)\neq g(j)}. \end{gathered} \tag{4.5}$$

This decomposition is surprisingly powerful, in part because models of Definition 2 asserts that d_j^w and d_j^b are statistically independent for any fixed group membership $g(1), g(2), \ldots, g(n)$. After separating the systematic bias term b in modularity from its random variation, we obtain the following decomposition.

Theorem 4.2.1 (Bias–variance decomposition for modularity)**.** *Under the null model of Definition 2, for b defined as in Eq.* (4.4) *and for a fixed (i.e., non-random) community assignment $g(1), g(2), \ldots, g(n)$, it holds that*

$$\widehat{Q} - b = \sum_{j=1}^{n} \alpha_j \left[d_j^w - \mathbb{E}\, d_j^w \right] + \sum_{j=1}^{n} \beta_j \left[d_j^b - \mathbb{E}\, d_j^b \right] + \mathcal{O}_P(\epsilon),$$

where the non-random quantities $\alpha_j = 1/2 + \beta_j$, β_j, and ϵ are defined as follows:

$$\beta_j = \left[\frac{1}{2} \frac{\sum_{l=1}^{n} \mathbb{E}\, d_l^w}{\mathbb{E}\|\boldsymbol{d}\|_1} - \frac{\mathbb{E}\, d_j^w}{\mathbb{E}\, d_j} \right], \quad 1 \leq j \leq n, \tag{4.6}$$

$$\epsilon = \frac{\sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)}}{\min(n, \|\boldsymbol{\pi}\|_1)\, \min_l \sqrt{\mathbb{E}\, d_l}}. \tag{4.7}$$

Proof. Since $\widehat{\mathbb{E}\, A}_{ij} = d_i d_j / \|\boldsymbol{d}\|_1$, modularity can be written as

$$\widehat{Q} = \sum_{j=1}^{n} \sum_{i<j} A_{ij} \delta_{g(i)=g(j)} - \sum_{j=1}^{n} \sum_{i<j} \widehat{\mathbb{E}\, A}_{ij} \delta_{g(i)=g(j)}. \tag{4.8}$$

We will show this theorem in six steps. We

1. Write $\widehat{\mathbb{E}\, A}_{ij}$ in terms of $\hat{\pi}_j = d_j / \sqrt{\|\boldsymbol{d}\|_1}$;

2. Expand the denominator $\sqrt{\|\boldsymbol{d}\|_1}$ around its mean in a convergent Taylor series;
3. Substitute $d_j = \mathbb{E}\, d_j + \mathcal{O}_P\big(\sqrt{\mathbb{E}\, d_j}\big)$ into the lower-order terms of the Taylor expansion of Step 2;
4. Apply the decomposition $d_j = d_j^w + d_j^b$, and center d_j^w and d_j^b about their respective means $\mathbb{E}\, d_j^w$ and $\mathbb{E}\, d_j^b$;
5. Collect all higher-order non-random terms in $\widehat{Q}$ into b;
6. Change the coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to add interpretability.

Let us first note some preliminaries. Denoting,

$$\|\boldsymbol{\pi}\|_1^{g(j),j} = \sum_{i\neq j} \pi_i\, \delta_{g(i)=g(j)} \quad \text{and} \quad \|\boldsymbol{\pi}\|_1^{\neg g(j)} = \sum_{i=1}^{n} \pi_i\, \delta_{g(i)\neq g(j)}, \tag{4.9}$$

we obtain for the expectations of d_j^w and d_j^b in Eq. (4.5)

$$\mathbb{E}\, d_j^w = \pi_j \|\boldsymbol{\pi}\|_1^{g(j),j} \quad \text{and} \quad \mathbb{E}\, d_j^b = \pi_j \|\boldsymbol{\pi}\|_1^{\neg g(j)}. \tag{4.10}$$

We are now ready to proceed with our analysis.

Step 1: We show in Lemma C.1.2 in Appendix C.1 that under Assumptions 1–4, it holds with ϵ as in Eq. (4.7) that

$$\widehat{Q} = \frac{1}{2}\sum_{j=1}^{n} d_j^w + \sum_{j=1}^{n}\sum_{i<j} \pi_i\pi_j \delta_{g(i)=g(j)} - \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon). \tag{4.11}$$

The main two steps are first, to recall from Eq. (3.20) that we may write

$$\widehat{\mathbb{E}\, A_{ij}} = \pi_i\pi_j + \pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j) + (\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j),$$

and second, to control the occurring error term using the result from Lemma B.1.7 in Appendix B.1.4. Namely, given Assumptions 1, 2, and 4, it holds that

$$\frac{(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}{\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)} = \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}}\right).$$

Step 2: In this step we focus on the penultimate term in Eq. (4.11). We appeal to a Taylor expansion of $(\|\boldsymbol{d}\|_1/\,\mathbb{E}\|\boldsymbol{d}\|_1)^{-1/2} = f(x) = x^{-1/2}$ at 1, and then control the remainder using Chebyshev's inequality. As a consequence, we obtain from Assumption 4 ($\operatorname{Var} A_{ij} = \Theta(\mathbb{E}\, A_{ij})$) that

$$\begin{aligned} &\sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} \\ &= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \cdot \left[1 - \frac{1}{2}\left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1\right) + \mathcal{O}_P\left(\frac{1}{\mathbb{E}\|\boldsymbol{d}\|_1}\right)\right]. \end{aligned} \tag{4.12}$$

We show in Lemma C.1.4 in Appendix C.1 that under Assumptions 1–4 it holds that

$$\sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \cdot \frac{1}{\mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}_P(\epsilon). \tag{4.13}$$

Continuing Eq. (4.12), we have that

$$= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1 \right) + \mathcal{O}_P(\epsilon). \tag{4.14}$$

Step 3: From Chebyshev's inequality and Assumption 4, we may conclude that $d_j = \mathbb{E}\, d_j \big[1 + \mathcal{O}_P\big(1/\sqrt{\mathbb{E}\, d_j}\big)\big]$. Inserting this result into the second (i.e., lower-order) term of the Taylor expansion in Eq. (4.14), we obtain

$$\begin{aligned} &= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \\ &\quad - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1 \right) \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_j}} \right) \right] + \mathcal{O}_P(\epsilon). \end{aligned} \tag{4.15}$$

Applying Chebyshev's inequality and then under Assumptions 1–4, it follows that

$$\begin{aligned} &\frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1 \right) \frac{1}{\sqrt{\mathbb{E}\, d_j}} \\ &= \mathcal{O}_P(\epsilon). \quad \text{(Lemma C.1.4 in Appendix C.1)} \end{aligned} \tag{4.16}$$

Applying Eq. (4.16) and then substituting $\sum_{j=1}^{n} d_j$ for $\|\boldsymbol{d}\|_1$ in Eq. (4.15), we obtain

$$\begin{aligned} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} &= \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon) \\ &\quad - \sum_{j=1}^{n} \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} \right] \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}. \end{aligned} \tag{4.17}$$

Step 4: Applying $d_j = d_j^w + d_j^b$ leads to the identity

$$\begin{aligned} &= \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon) \\ &\quad - \sum_{j=1}^{n} \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} \right] \frac{d_j^w}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \\ &\quad - \sum_{j=1}^{n} \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} \right] \frac{d_j^b}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}. \end{aligned} \tag{4.18}$$

We define non-random factors $\alpha_j^* = 1/2 + \beta_j^*$ and β_j^* as

$$\beta_j^* = \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} \right] \frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}. \tag{4.19}$$

In Lemma C.1.5 in Appendix C.1, we address that α_j^* and β_j^* defined here differ from α_j and β_j defined in Theorem 4.2.1. Combining the results from Eqs. (4.11) and (4.18), we may rewrite $\widehat{Q}$ in terms of α_j^* and β_j^* as

$$\widehat{Q} = \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} + \sum_{j=1}^{n} \alpha_j^* d_j^w + \sum_{j=1}^{n} \beta_j^* d_j^b + \mathcal{O}_P(\epsilon).$$

After centering d_j^w and d_j^b about their respective means, we obtain

$$\begin{aligned} \widehat{Q} = &\sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right] + \sum_{j=1}^{n} \alpha_j^* \mathbb{E}\, d_j^w + \sum_{j=1}^{n} \beta_j^* \mathbb{E}\, d_j^b \\ &+ \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon). \end{aligned} \tag{4.20}$$

Step 5 We show in Lemma C.1.3 in Appendix C.1 that under Assumptions 1–4, we may combine all non-random terms in modularity to $b + \mathcal{O}(\epsilon)$ and in turn obtain

$$\widehat{Q} = \sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] - \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right] + b + \mathcal{O}(\epsilon). \tag{4.21}$$

In the proof of Lemma C.1.3, we show that the non-random terms in Eq. (4.20):

a) $\sum_{j=1}^{n} \alpha_j^* \mathbb{E}\, d_j^w + \sum_{j=1}^{n} \beta_j^* \mathbb{E}\, d_j^b$,

b) $\sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}$

are equal via straightforward algebraic computations. A Taylor expansion, followed by more algebraic computations, and an upper-bound on the lower order terms leads to

$$a) + b) = \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \left[\frac{\pi_i + \pi_j}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right] \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon).$$

In order to gain interpretability, we rearrange the term $a) + b)$ even further:

$$= \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} \left(\mathbb{E}\, d_i + \mathbb{E}\, d_j - \|\boldsymbol{\pi}\|_2^2\right)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon). \tag{4.22}$$

This leads to the result of Lemma C.1.3 that all non-random terms in modularity in Eq. (4.20) may be summed to $b + \mathcal{O}(\epsilon)$. Inserting the results from Eqs. (4.4) and (4.22) into Eq. (4.20) and under the assumption that all error terms are controlled (see Lemma C.1.4 in Appendix C.1), we obtain

$$\widehat{Q} = \sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] - \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right] + b + \mathcal{O}(\epsilon). \tag{4.23}$$

Step 6: To add interpretability to the coefficients $\boldsymbol{\alpha}^* = 0.5 + \boldsymbol{\beta}^*$ and $\boldsymbol{\beta}^*$, we change the formulation from β_j^* in Eq. (4.19) to β_j in Theorem 4.2.1 (see Eq. (4.6)). By doing so, we add an error term into the decomposition of modularity that asymptotically wears off. For more details see Lemma C.1.5.

As a consequence, we conclude the required result of Theorem 4.2.1; i.e.,

$$\widehat{Q} - b = \sum_{j=1}^{n} \alpha_j \left[d_j^w - \mathbb{E}\, d_j^w \right] + \sum_{j=1}^{n} \beta_j \left[d_j^b - \mathbb{E}\, d_j^b \right] + \mathcal{O}_P(\epsilon).$$

□

Theorem 4.2.1 quantifies the random variability inherent in modularity under a model of Definition 2. It establishes that a main term contributing to the variability of $\widehat{Q} - b$ in this setting is a linear combination of centered within- and between-group degrees (d_j^w, d_j^b), which for each j are statistically independent. The weights α_j and β_j associated with this linear combination are determined by the global proportion of expected within-group edges in the network, relative to the local proportion of expected within-group edges specific to node j.

4.2.2 A limit theorem for modularity

Our main result in this chapter is a methodology to understand objectively whether a covariate captures the structure of the interactions in a network. Technically, we derive a theorem quantifying the large-sample behavior of modularity in the setting above. In particular, if a null model of Definition 2 is in force, then modularity in the presence of covariates behaves like a Normal random variable for large networks. This enables us to associate a p-value with any observed community structure, quantifying how unlikely it is (under the null) to observe a community structure *at least as extreme as* the one we observe.

To derive the asymptotic distribution of modularity, we combine the two insights from the previous section: we shift modularity $\widehat{Q}$ by its approximate bias b in Eq. (4.4) and then scale it by the standard deviation s of $\sum_{i=1}^{n} \alpha_i [d_i^w - \mathbb{E}\, d_i^w] + \sum_{i=1}^{n} \beta_i \left[d_i^b - \mathbb{E}\, d_i^b \right]$:

$$s^2 = \sum_{j=1}^{n} \sum_{i<j} \left[\delta_{g(i)=g(j)} + \beta_i + \beta_j \right]^2 \operatorname{Var} A_{ij}. \tag{4.24}$$

Recalling Theorem 4.2.1, we then know that we are left with a linear combination of centered within- and between-group degrees that are now also scaled by s. This leads to the following central limit theorem for modularity $\widehat{Q}$.

Theorem 4.2.2 (Central limit theorem for modularity)**.** *Suppose a null model of Definition 2 is in force, and consider a sequence of networks where for each n we observe a fixed (non-random) group membership $g(1), g(2), \ldots, g(n)$. Then as long as the number K of communities grows strictly more slowly than n (i.e., $K/n \to 0$), and as $n \to \infty$,*

$$\frac{\widehat{Q} - b}{s} \xrightarrow{d} \mathrm{Normal}(0, 1),$$

with b and s as defined in Eqs. (4.4) *and* (4.24)*.*

Proof. For convenience, we use β_j^* (Eq. (4.19)) in this proof instead of β_j (Eq. (4.6)). We have seen in the proof of Theorem 4.2.1 that β_j and β_j^* are asymptotically equivalent; i.e., $\beta_j^* = \beta_j[1 + \mathcal{O}(1/n)]$ (see Eq. (C.22)) and that substituting one for the other in the bias-variance decomposition asymptotically wears off (see Lemma C.1.5).

Recalling the definitions of $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ from Eq. (4.19), we define a sequence of random variables:

$$X_n = \sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j^* \Big[d_j^b - \mathbb{E}\, d_j^b\Big]. \tag{4.25}$$

In Lemma C.1.6 in Appendix C.1.3 we show that we may write X_n as a sum of independent, zero-mean random variables:

$$\begin{aligned} X_n &= \sum_{j=1}^{n} \sum_{i<j} c_{ij}[A_{ij} - \mathbb{E}\, A_{ij}], \\ c_{ij} &= \delta_{g(i)=g(j)} + \beta_i^* + \beta_j^*. \end{aligned} \tag{4.26}$$

Furthermore, we show in Lemma C.1.7 in Appendix C.1.3 that $c_{ij} = \mathcal{O}(1)$ and that it may be expressed as a function only of the group assignments $g(i) = m$ and $g(j) = l$:

$$c_{lm} = \delta_{l=m} + \sum_{k=1}^{K} \left(\frac{\|\boldsymbol{\pi}\|_1^{k,\emptyset}}{\|\boldsymbol{\pi}\|_1}\right)^2 - \frac{\|\boldsymbol{\pi}\|_1^{l,\emptyset}}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_1^{m,\emptyset}}{\|\boldsymbol{\pi}\|_1} + \mathcal{O}\left(\frac{1}{n}\right).$$

For the remaining proof, we first show that $(\mathrm{Var}\, X_n)^{-\frac{1}{2}} X_n \xrightarrow{d} \mathrm{Normal}(0, 1)$ using the Lindeberg–Feller Central Limit Theorem (see Appendix A.2) and second, we control the discrepancy between $(\mathrm{Var}\, X_n)^{-\frac{1}{2}} X_n$ and $\Big(\widehat{Q} - b\Big)/s$. For the Lindeberg–Feller Central Limit Theorem, we need to show that the following two sufficient conditions are fulfilled.

1. $\mathrm{Var}(c_{ij} A_{ij}) < \infty$;

2. The Lyapunov condition for exponent 1 is satisfied; i.e.,

$$\frac{\sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\Big[(c_{ij} A_{ij} - \mathbb{E}(c_{ij} A_{ij}))^3\Big]}{\Big[\mathrm{Var}\Big(\sum_{j=1}^{n} \sum_{i<j} c_{ij} A_{ij}\Big)\Big]^{3/2}} \to 0.$$

Condition 1:

$$\begin{aligned}\operatorname{Var}(c_{ij}A_{ij}) &= c_{ij}^2 \operatorname{Var}(A_{ij}) \\ &= c_{ij}^2 \Theta(\pi_i \pi_j) \quad \text{(Assumption 4)} \\ &< \infty. \quad \text{(Lemma C.1.7: } c_{ij} = \mathcal{O}(1);\ \pi_i, \pi_j \in \mathbb{R}_{>0})\end{aligned}$$

Condition 2:

$$\begin{aligned}&\frac{\sum_{j=1}^n \sum_{i<j} \mathbb{E}\Big[(c_{ij}A_{ij} - \mathbb{E}(c_{ij}A_{ij}))^3\Big]}{\Big[\operatorname{Var}\Big(\sum_{j=1}^n \sum_{i<j} c_{ij}A_{ij}\Big)\Big]^{3/2}} \\ &= \frac{\sum_{j=1}^n \sum_{i<j} c_{ij}^3 \mathbb{E}\Big[(A_{ij} - \mathbb{E}(A_{ij}))^3\Big]}{\Big[\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij}\Big]^{3/2}} \\ &= \mathcal{O}(1) \cdot \frac{\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \mathbb{E}\Big[(A_{ij} - \mathbb{E}(A_{ij}))^3\Big]}{\Big[\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij}\Big]^{3/2}} \quad \text{(Lemma C.1.7: } c_{ij} = \mathcal{O}(1)) \\ &= \mathcal{O}\left(\frac{\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij}}{\Big[\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij}\Big]^{3/2}}\right) \quad \text{(Assumption 5)} \\ &= \mathcal{O}\left(\frac{1}{\Big[\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij}\Big]^{1/2}}\right). \quad \text{(Lemma C.1.7: } c_{ij} = \mathcal{O}(1))\end{aligned}$$

For Condition 2, it remains to show that $\sum_{j=1}^n \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij} \to \infty$. We substitute

$$a_k = \|\boldsymbol{\pi}\|_1^{k,\emptyset}, \tag{4.27}$$

such that $\|\boldsymbol{a}\|_1 = \|\boldsymbol{\pi}\|_1$. In Lemma C.1.9, we show that under Assumptions 1 and 4, and whenever $K = o(n)$, it holds that

$$\sum_{j=1}^{n} \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij} = \Theta\Big(\|\boldsymbol{a}\|_2^2\Big).$$

Now, since $\|\boldsymbol{\pi}\|_1 \to \infty$ (Assumption 2), and by construction $\|\boldsymbol{a}\|_1 = \|\boldsymbol{\pi}\|_1$, we see

$$\begin{aligned}\|\boldsymbol{a}\|_2^2 &\geq \frac{\|\boldsymbol{\pi}\|_1^2}{K} \quad \Big(K\|\boldsymbol{a}\|_2^2 \geq \|\boldsymbol{a}\|_1^2\Big) \\ &= \omega\Big(\frac{n}{K}\Big) \quad \text{(Assumption 2)} \\ &= \omega(1). \quad (K = o(n))\end{aligned}$$

Thus the Lyapunov condition is satisfied, and via the Lindeberg–Feller Central Limit Theorem (see Appendix A.2) we obtain the claimed result that

$$(\operatorname{Var} X_n)^{-\frac{1}{2}} X_n \xrightarrow{d} \operatorname{Normal}(0,1). \tag{4.28}$$

Combining Theorem 4.2.1 and Eq. (4.25), we obtain that modularity $\widehat{Q}$ satisfies

$$\begin{aligned} &\widehat{Q} = b + X_n + \mathcal{O}_p(\epsilon) \\ \Rightarrow \quad &(\operatorname{Var} X_n)^{-\frac{1}{2}} \Big(\widehat{Q} - b\Big) = (\operatorname{Var} X_n)^{-\frac{1}{2}} X_n + (\operatorname{Var} X_n)^{-\frac{1}{2}} \mathcal{O}_p(\epsilon). \end{aligned} \tag{4.29}$$

We show in Lemma C.1.8 in Appendix C.1.3 that under Assumptions 1–3

$$(\operatorname{Var} X_n)^{-\frac{1}{2}} \epsilon \xrightarrow{n} 0.$$

We are now ready to complete the proof of Theorem 4.2.2. Since β and β^* are asymptotically equivalent (see Eq. (C.22)), we observe from Eq. (4.26) that s as defined in Theorem 4.2.2 (see Eq. (4.24)) satisfies

$$s^2 = \operatorname{Var} X_n.$$

Combining the results from Eqs. (4.28), (4.29), and Lemma C.1.8 using Slutsky's Theorem (see Appendix A.2), we conclude the overall result of this theorem; i.e.,

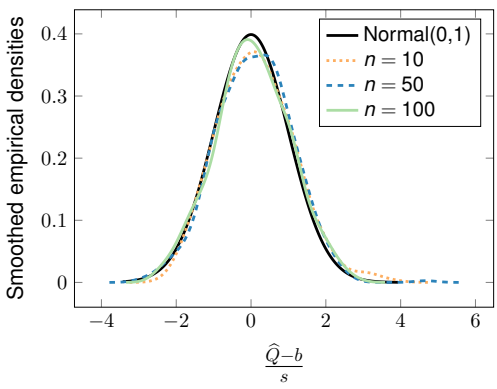
$$\frac{\widehat{Q} - b}{s} \xrightarrow{d} \operatorname{Normal}(0,1).$$

□

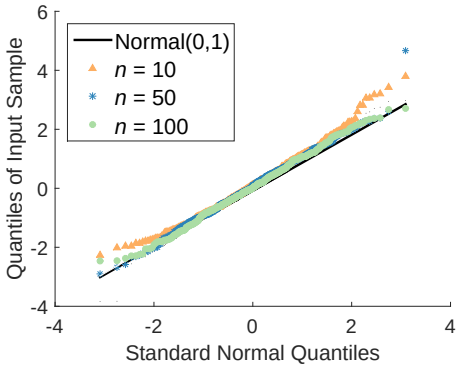
We now completed the proof of Theorem 4.2.2 that shows if a model of Definition 2 is in force, then modularity in the presence of covariates behaves like a Normal random variable. This enables us to associate a p-value with any observed community structure, quantifying how unlikely it is (under the null) to observe a community structure *at least as extreme as* the one we observe.

4.3 Illustrative simulations for the limit theorem for modularity

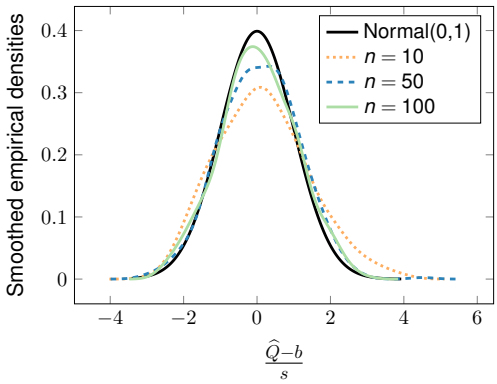
We now illustrate the theoretical result of Theorem 4.2.2 on the large-sample distribution of modularity. To emphasize that the broad class of the degree-based models of Definition 2 is nonparametric, we simulate data from networks following several different distributions: simple networks with Bernoulli distributed edges (see Section 4.3.1); and networks with multiple edges modeled as Binomial, Poisson, or Negative Binomial distributed (Section 4.3.2). The simulations include all models used for the data analysis in Chapter 5.



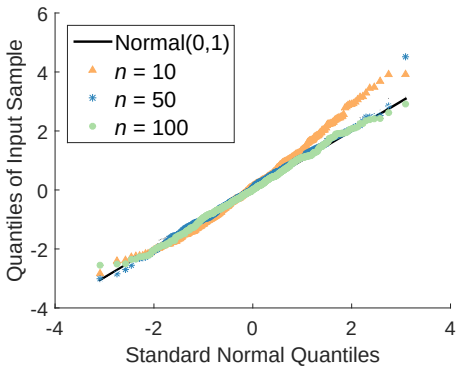
(a) Smoothed empirical densities



(b) Q-Q plot



(c) Plug-in: Smoothed empirical densities



(d) Plug-in: Q-Q plot

Figure 4.2: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$ for simple networks; simulated from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.6}$.

4.3.1 Simple networks

For simple random graphs, we model the edges as Bernoulli random variables with success probability $\pi_i \pi_j$. In agreement with the simulations in Chapter 3, we generate the parameters π_i as $\pi_i = \theta\ i^{-\gamma}$, for $i = 1, \ldots, n$ with $1/2 < \gamma < 1$ to match the power law degree behavior often observed in practice [48, p. 11]. We generate a random community assignment into four communities. We then draw 500 independent realizations from this model and estimate for each of these networks modularity $\widehat{Q}$ as defined in Eq. (4.1):

$$\widehat{Q} = \sum_{j=1}^{n} \sum_{i<j} \left[A_{ij} - \frac{d_i d_j}{\|\boldsymbol{d}\|_1} \right] \delta_{g(i)=g(j)}.$$

We then center and scale modularity by its approximated bias b in Eq. (4.4), and standard deviation s in Eq. (4.24) leading to

$$\frac{\widehat{Q} - b}{s}.$$

As for the simulations in Chapter 3, we look at the smoothed empirical density and the quantile-quantile plots (Q-Q plot) of the 500 simulated samples. We repeat the procedure increasing the number of nodes n.

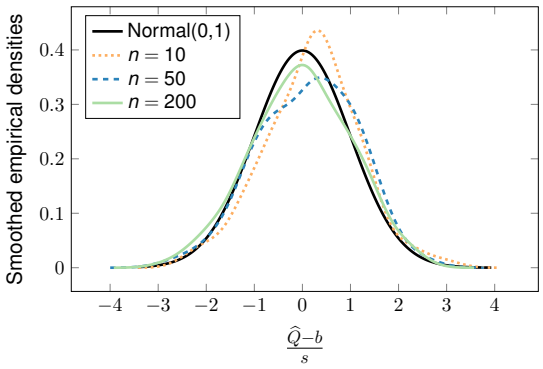
Figures 4.2a and 4.2b illustrate the results for $\pi_i = i^{-0.6}$. We see that already for a small number of nodes ($n = 10$) the approximation of the distribution of modularity by a $\mathrm{Normal}(0, 1)$ performs well. However, since in practice, we cannot observe the parameters $\pi_1, \ldots, \pi_n$ we run 500 simulations standardizing modularity with a plug-in estimator for b and s where we substitute all π_is by $\hat{\pi}_i$s. In Figures 4.2c and 4.2d, we display the results observing that for small n (e.g. $n = 10$) the normal approximation of the distribution of modularity with a plug-in standardization is miserable but that as n increases it improves. We notice that here the convergence in distribution of modularity seems to be much faster than the convergence in distribution for the standardized $\hat{\pi}_5$ in Figure 3.1, where in both cases we simulate from $\pi_i = i^{-0.6}$.

In Appendix C.2.1, we show a representative selection of the results for simulations with different levels of sparsity, where we vary $\theta \in (0, 1]$, and $\gamma \in (0, 1)$ (e.g., Figures C.1: $\pi_i = i^{-0.2}$, C.2: $\pi_i = 0.9\ i^{-0.7}$). In addition, we illustrate that Theorem 4.2.2 also holds for Erdős-Rényi graphs as a special case of a degree-based model (Figure C.3: $\pi_i = 0.4$). In all cases, we observe that as n increases the difference between the smoothed empirical density of modularity $\widehat{Q}$ and a $\mathrm{Normal}(0, 1)$ density shrinks. We see further that when the network is more dense the same number of nodes leads to a smaller difference between the smoothed empirical density and a $\mathrm{Normal}(0, 1)$. Comparing the simulations from the same model between Chapters 3 and 4, we get the impression that for the same number of nodes the empirical distribution of modularity $\widehat{Q}$ is closer to a $\mathrm{Normal}(0, 1)$ distribution than the empirical distribution of $\hat{\pi}_i$.

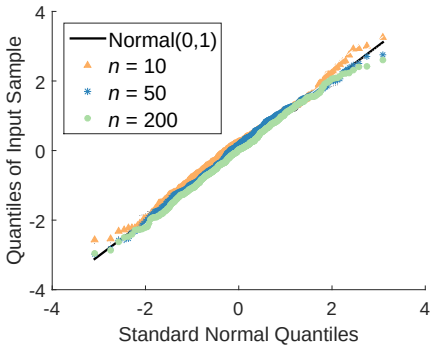
4.3.2 Multi-edge networks

Motivated by the preceding data analysis (see Chapter 5), we here simulate networks with multiple edges from a degree-based model. In a consecutive order, we model the edges to follow a Binomial, Poisson, and Negative Binomial distribution. In all cases, we simulate $\mathbb{E}\, A_{ij} = \theta^2 (ij)^{-\gamma}$ as before, but for $\theta \in \mathbb{N}_{>0}$.

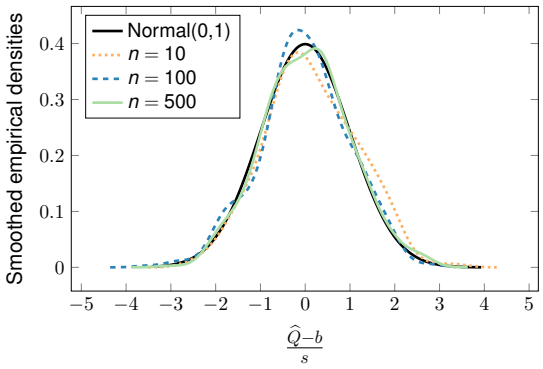
In Figure 4.3, we observe for all three edge distributions that as n increases the difference between the smoothed empirical density of modularity $\widehat{Q}$ and a $\mathrm{Normal}(0,1)$ density shrinks in agreement with Theorem 4.2.2. For each of the distributions, we simulate 500 samples from



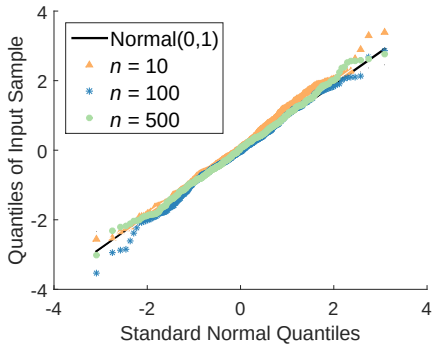
(a) Binomial: Smoothed empirical densities



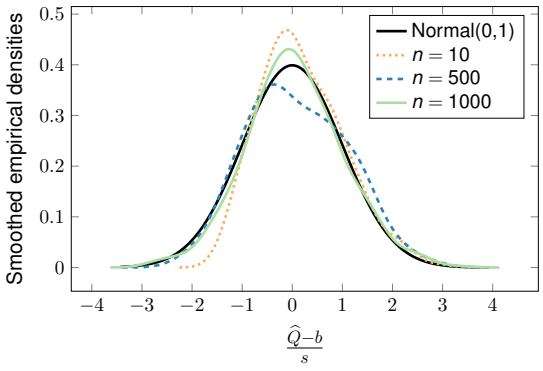
(b) Binomial: Q-Q plot



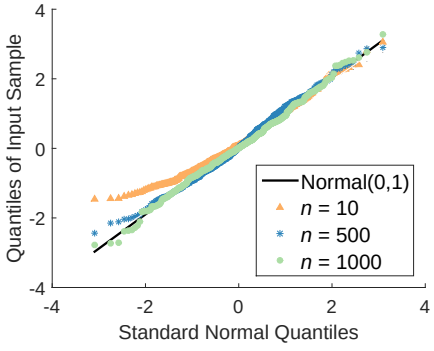
(c) Poisson: Smoothed empirical densities



(d) Poisson: Q-Q plot



(e) Negative Binomial: Smoothed empirical densities



(f) Negative Binomial: Q-Q plot

Figure 4.3: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$ for multi-edge networks standardized by the plug-in estimators for b and s; simulated from power law networks with $\mathbb{E}\, A_{ij} = 11.56 \cdot (ij)^{-0.6}$.

a degree-based model with $\mathbb{E}\, A_{ij} = 11.56 \cdot (ij)^{-0.6}$. Each time, we compute the empirical modularity $\widehat{Q}$ (see Eq. (4.1)) and standardize it using a plug-in estimator for b and s where we substitute all π_is by $\hat{\pi}_i$s. Since we fix each of the edge expectations $\mathbb{E}\, A_{ij}$ for a given n (such that they are equal for all three distributions), we ensure that the networks are equally sparse. Comparing the simulations of the three distributions, we observe that for the same number of nodes the distance between the empirical distribution and a $\mathrm{Normal}(0,1)$ is larger for the Poisson and Negative Binomial distributed edges—the skewed distributions—than for the Binomial distributed edges.

In more detail, to simulate from a *Binomial distribution*, we assume that there are $M = 13$ (figure chosen at random) potential edges between every pair of nodes, and model the network:

$$\begin{aligned} A_{ij} &\sim \mathrm{Binomial}\big(M, \theta^2 (ij)^{-\gamma}/M\big), \\ \Rightarrow \mathrm{Var}\, A_{ij} &= \mathbb{E}\, A_{ij}(1 - \mathbb{E}\, A_{ij}/M). \end{aligned}$$

Then, to simulate from a *Poisson distribution*, we model the network:

$$\begin{aligned} A_{ij} &\sim \mathrm{Poisson}(\pi_i \pi_j), \\ \Rightarrow \mathrm{Var}\, A_{ij} &= \mathbb{E}\, A_{ij}. \end{aligned}$$

To simulate from a *Negative Binomial distribution* ($\mathrm{NB}(\boldsymbol{\mu}, r)$) we assume the shape parameter $r = 0.24$ (figure chosen at random) and with Γ denoting the gamma function we obtain the probability distribution function:

$$\begin{aligned} f(A_{ij}|\mu_{ij}, r) &= \frac{\Gamma(A_{ij} + r)}{\Gamma(A_{ij} + 1)\Gamma(r)} \left(\frac{r}{\mathbb{E}\, A_{ij} + r}\right)^r \left(\frac{\mathbb{E}\, A_{ij}}{\mathbb{E}\, A_{ij} + r}\right)^{A_{ij}}, \\ &\Rightarrow \mathrm{Var}\, A_{ij} = \mathbb{E}\, A_{ij} + (\mathbb{E}\, A_{ij})^2/r. \end{aligned}$$

For better interpretation, this is equivalent to modeling the edges as Poisson distributed with mean Z where we regard the mean itself as gamma distributed with $\mathbb{E}\, Z = \mu_{ij}$ and shape parameter r [96, p. 199]. Under this parametrization we obtain an expectation $\mathbb{E}\, A_{ij} = \mu_{ij}$, and a skewness $\gamma_{ij} = (2\mu_{ij} + r)/\sqrt{(\mu_{ij} + r)\mu_{ij} r}$ [73, p. 216].

In Figure C.4 in Appendix C.2.2, we observe that for all three distributions when the network is less sparse ($i.e., \mathbb{E}\, A_{ij} = 11.56 \cdot (ij)^{-0.2}$) the distance between the empirical distribution and a $\mathrm{Normal}(0,1)$ is smaller for the same number of nodes. This once more illustrates that the convergence in distribution in Theorem 4.2.2 depends on the effective sample size rather than the absolute number of nodes n.

4.4 Discussion

In this chapter, we delivered theoretical results to assess the significance of observed community structure, enabling us to identify informative covariate-based community assignments while controlling the Type I error. To do so, we derived for the first time a theoretical foundation for modularity [106]—a popular method for community detection—and extended its applicability to objectively assess the strength of observed community structures. Our approach built up on the work by Arias-Castro and Verzelen [8] who utilized a simplified version of modularity to detect the presence of network community structure. After the posting of [56] describing the work of this thesis, Newman [109] derives a complementary interpretation of modularity: community detection using a generalized modularity (see Eq. (2.3), [123]) is closely related to maximizing the likelihood of the degree-corrected stochastic blockmodel.

In technical terms, we have established a central limit theorem for modularity under a nonparametric null model, yielding p-values to assess the significance of observed community structure. The model we introduced shows explicitly how modularity measures variability in the data that cannot be explained solely by node-specific propensities for connection. What is more, modularity has more explanatory power than a classical (chi-squared) goodness-of-fit statistic: by aggregating the estimated *signed* residuals $A_{ij} - d_i d_j / \|\boldsymbol{d}\|_1$ within every network community, it measures the global tendency of a given community assignment to explain the observed network structure.

Chapter 5

Data analysis

After establishing the asymptotic behavior of modularity in the presence of covariates in Chapter 4, we now show how to apply these results in practice. We first explain a step-wise procedure how to turn the theoretical results into a methodology (see Section 5.1). Second, we validate our methodology on benchmark examples where the reported covariate has previously been used as ground truth for community detection (see Section 5.2). Third, we evaluate communities in a multi-edge network (see Section 5.3) and finish with a discussion about the data analysis (see Section 5.4).

5.1 A methodology to quantify network structure

To turn our theory into a methodology suitable for a specific network dataset, we first need to elicit a model for the data based on Definition 2. We then fit this model, leading ultimately to a p-value based on Theorem 4.2.2. We now explain the corresponding four-step procedure.

1. First, we must further specify the null model of Definition 2, so that the parameter s^2 in Eq. (4.24) can be estimated. This can be done either by assuming sets of the variances $\operatorname{Var} A_{ij}$ to be equal, or by assuming a distribution for the edges A_{ij}. For instance, since the benchmark networks we consider in Section 5.2 are binary ($A_{ij} \in \{0, 1\}$), we model their edges as $A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j)$.

2. Second, we must assess whether the five asymptotic assumptions of Definition 2 appear to hold for our data and whether the number K of communities is sufficiently smaller than n (i.e., we assume $K/n \to 0$). Technical Assumptions 4 ($\operatorname{Var} A_{ij} / \mathbb{E}\, A_{ij} = \Theta(1), \forall i, j$) and 5 $\left(\mathbb{E}\left[(A_{ij} - \mathbb{E}\, A_{ij})^3\right] / \operatorname{Var}(A_{ij}) = \mathcal{O}(1), \forall i, j\right)$ exclude extreme behavior, and are therefore for many distributions automatically satisfied. For instance, both are fulfilled whenever $A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j)$ or $A_{ij} \sim \text{Poisson}(\pi_i \pi_j)$. Assumption 3 controls the

growth of $\mathbb{E}\, A_{ij}$ with n, and only needs to be verified for weighted or multi-edge networks. Since we cannot observe π_i we assess Assumptions 1–3 by substituting $\hat{\pi}_i$ for π_i, noting that

$$\begin{aligned}
&\text{Assumption 1: } \max_i \pi_i/\bar{\pi} = \mathcal{O}(1) && \Rightarrow \max_i \hat{\pi}_i/\bar{\hat{\pi}} = \max_i d_i/\bar{d},\\
&\text{Assumption 2: } \min_i \pi_i \cdot \sqrt{n} = \omega(1) && \Rightarrow \min_i \hat{\pi}_i \cdot \sqrt{n} = \min_i d_i \Big/ \sqrt{\bar{d}},\\
&\text{Assumption 3: } \max_i \pi_i \big/ \sqrt{n} = o(1) && \Rightarrow \max_i \hat{\pi}_i \big/ \sqrt{n} = \max_i d_i \Big/ \Big(n\sqrt{\bar{d}}\Big).
\end{aligned}$$

Replacing $\min_i d_i$, $\bar{d}$, and $\max_i d_i$ by the first, second and third degree quartiles, respectively enables us to quantify the assumptions in a robust way.

3. Third, we estimate the parameters b and s necessary to shift and scale $\widehat{Q}$ in accordance with Theorem 4.2.2. To obtain an estimator $\hat{b}$, we substitute $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ in Eq. (4.4). The estimator $\hat{s}$ depends on the assumption added in Step 1 above. For instance, when $A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j)$, we have $\text{Var}\, A_{ij} = \pi_i\pi_j(1 - \pi_i\pi_j)$. Then, $\hat{s}$ follows directly by substituting $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ in Eq. (4.24).

4. Finally, we compute and interpret the resulting approximate p-value. We first define a community assignment $g(1), g(2), \ldots, g(n)$ based on a covariate, and calculate $\widehat{Q}$ as per Definition 1. We next estimate $(\widehat{Q} - b)/s$ using $\hat{b}$ and $\hat{s}$. Then, by Theorem 4.2.2, we compute an approximate one-sided p-value as follows:

$$\Pr\left(Z \geq \frac{\widehat{Q} - \hat{b}}{\hat{s}}\right), \quad Z \sim \text{Normal}(0, 1). \tag{5.1}$$

A small p-value implies that the observed value of modularity (or any larger value) is unlikely under the null model of Definition 2.

5.2 Validating the methodology on benchmark examples

We now illustrate the complete analysis procedure for four binary networks which, along with their covariates, frequently serve as benchmarks for community detection [45, 108]. Table 5.1 and 5.2 summarize the data and the results, respectively.

5.2.1 Description of the data

Table 5.1 summarizes the four benchmark networks that we analyze in this section. All four are binary networks that reflect social interactions of different types; varying in their number of

nodes between 105 and 36297.

The first example is a network of books about U.S. politics where two books are connected if they have frequently been purchased together from the on-line bookseller Amazon.com [108]. The 105 books are categorized by their stated or apparent political alignment into conservative, liberal, and neutral. In Figure 4.1, we use this dataset to illustrate the decomposition of a network into within- and between-group edges.

The second example is a network of 198 jazz bands that performed between 1912 and 1940 [59]. Two bands are connected if they have at least one band member in common. The bands are categorized by their 17 different recording locations; where New York and Chicago dominate with being reported by about 45% and 34% of the bands, respectively.

The third example is a network of political commentary websites (weblogs) of a single day snapshot; collected by Adamic and Glance with particular interest in the 2004 U.S. presidential election [1]. The weblogs are connected if either of the corresponding weblogs contains a hyperlink to the other on the front page. Following Newman [108], we restrict our analysis to the largest component of 1224 weblogs; categorized by their political alignment into conservative and liberal.

The last benchmark example is a network of physicists where two researchers are connected if they have co-authored a manuscript between Jan 1, 1995 and December 31, 1999 on either the *Astrophysics-*, *Condensed Matter-* or *High-Energy Theory E-Print arXiv* [104]. We define the community assignment based on the arXiv categories: three groups for authors who published in either of the categories, three for groups of pairs of categories and one group for authors who published in all three categories; leading to 7 communities in total. We exclude researchers that are not connected; leading to 36297 nodes.

5.2.2 Elicitation of the model and deriving the p-values

Applying the four-step procedure described in Section 5.1, we validate our methodology using the four benchmark datasets; starting with elicitation of the model for the data.

1. Since the benchmark networks we consider here are binary ($A_{ij} \in \{0, 1\}$), we further specify the null model of Definition 2 by assuming:

$$A_{ij} \sim \text{Bernoulli}(\pi_i \pi_j).$$

2. We must assess whether the asymptotic assumptions of Definition 2 appear to hold. Assumptions 3–5 are automatically satisfied for Bernoulli edges, and since all $K \ll n$ we

Dataset	Covariate	Nodes	Groups	Degree percentiles		
				25%	50%	75%
Books [108]	Political alignment	105	3	5	6	9
Jazz bands [59]	Recording location	198	17	16	25	39
Weblogs [1]	Political alignment	1224	2	3	13	36
Co-authors [104]	Subject category	36297	7	2	5	10

Table 5.1: Validation of the model assumptions of Definition 2 for four benchmark network datasets.

Dataset: Covariate	Simulated under the null				Data as observed	
	$(\widehat{Q}-\hat{b})/\hat{s}$		p-value		$(\widehat{Q}-\hat{b})/\hat{s}$	p-value
	mean	std.	mean	std.		
Books: Pol. align. [108]	0.02	1.01	0.51	0.29	21	$< 10^{-6}$
Jazz bands: Rec. loc. [59]	0.01	1.02	0.51	0.29	29	$< 10^{-6}$
Weblogs: Pol. align. [1]	0.01	1.04	0.50	0.30	118	$< 10^{-6}$
Co-authors: Subj. cat. [104]	0.00	1.00	0.50	0.29	472	$< 10^{-6}$

Table 5.2: Analysis of the four benchmark network datasets from Table 5.1, using modularity derived from covariate-based community assignments.

are left to assess Assumptions 1 ($\max_i d_i/\bar{d}$ bounded) and 2 ($\min_i d_i\Big/\sqrt{\bar{d}}$ growing). Denoting Q_1, Q_2, and Q_3 the first, second and third degree quartiles, we assess Assumptions 1 and 2 by the ratios Q_3/Q_2 and $Q_1/\sqrt{Q_2}$, respectively. As shown in Table 5.1, we observe that for all four benchmark networks, these ratios are of order one. This indicates that these networks are neither too star-like nor too sparse for Theorem 4.2.2 to apply.

3. We estimate the parameters b and s necessary to shift and scale $\widehat{Q}$. From the assumption that $A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j)$ in Step 1, it follows that

$$\text{Var}\, A_{ij} = \pi_i\pi_j(1-\pi_i\pi_j).$$

Then, $\hat{b}$ and $\hat{s}$ follow directly by substituting $\hat{\boldsymbol{\pi}}$ for $\boldsymbol{\pi}$ in Eqs. (4.4) and (4.24).

4. Finally, we compute the resulting approximate p-values listed in Table 5.2. We define communities $g(1), g(2), \ldots, g(n)$ based on the corresponding covariate (see Table 5.1), and calculate: $\Pr\Big(Z \geq \frac{\widehat{Q}-\hat{b}}{\hat{s}}\Big), \quad Z \sim \text{Normal}(0,1)$.

5.2.3 Results

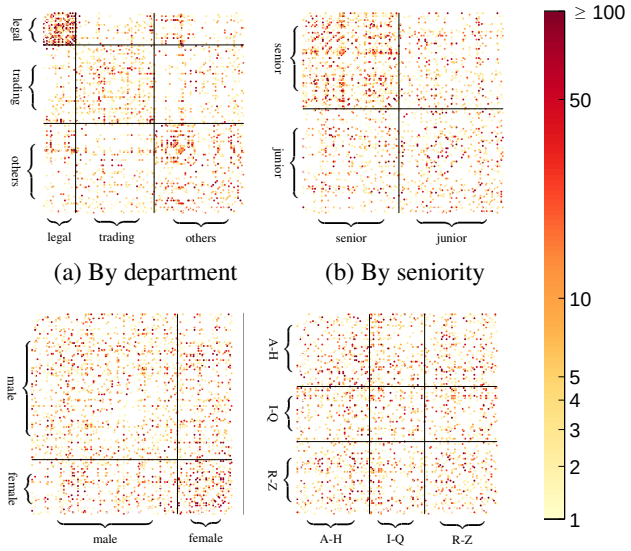
To validate the methodology introduced in this work, we address two key questions. First, we analyze the behavior of the p-value when the null model of Definition 2 that does not support community structure is in place. Second, we investigate the behavior of the p-value for observed networks, and their covariates which previously have been used as ground truth for community detection.

The first conclusion of our benchmark analysis is that when we simulate from the fitted model the empirical results agree with Theorem 4.2.2. First, we fit the null model of Definition 2 using maximum likelihood estimation to each of these four networks, and then simulate from the fitted model (parametric bootstrap). As a result we obtain 10000 simulated networks per dataset each following the null model. For each sample, we compute $(\widehat{Q} - \hat{b})/\hat{s}$ and the corresponding p-value (via Eq. (5.1)) for the respective covariate in Table 5.1. As a result, we see in Table 5.2 that each set of 10000 simulated networks results in a p-value with empirical mean near $1/2$ and standard deviation near $1/\sqrt{12}$. This empirical result aligns with Theorem 4.2.2, which predicts the p-values to be uniformly distributed with exactly that mean and standard deviation in the limit.

Our second conclusion is that, when using the observed data rather than simulated data under the null, each of the covariates leads to a very small p-value ($< 10^{-6}$; see Table 5.2). This suggests that the data as observed are extremely unlikely under the null model of Definition 2. Furthermore, since the null itself cannot explain any community structure, the conclusion we obtain agrees with the use of these covariates by other researchers as ground truth in community detection settings.

5.3 Evaluating communities in a multi-edge email network

We now illustrate how our methodology can identify covariates that reflect a network's community structure. This analysis goes beyond the four benchmark examples considered in Section 5.2, where we validated our methodology but did not reach any new data-analytic conclusions. Here we evaluate the effects of employee *seniority*, *gender*, and *company department* on community structure in a multi-edge corporate email network. Table 5.4 summarizes all results, showing that each of these covariates results in a small p-value, while covariates based on grouping the *first-* or *last-name initials* of the employees do not. We will return to this analysis in more detail below, after describing the data and eliciting a suitable model.



(a) By department (b) By seniority

(c) By gender (d) By last name initial

Figure 5.1: Multi-edges A_{ij} in the Enron corporate email dataset (153 employees, 32261 pairwise email exchanges), grouped according to four different covariate-based community assignments. Shading indicates the number of emails exchanged.

5.3.1 Description of the data

Figure 5.1 illustrates four community structures of a multi-edge email interaction network based on the employee seniority, gender, company department, and last name initial. This network and its covariates form a substantially richer dataset than those treated in Section 5.2. The data come from the Enron Corporation [33, 120]: as part of a U.S. government investigation following allegations of fraud, the email activities of employees from 1998–2002 were made public. Following the analysis in [120], we exclude all emails that have been sent *en masse* (to more than five recipients), leading to 32,261 pairwise email exchanges between 153 employees.

This corporate email network has previously been analyzed by Perry and Wolfe [120] where the authors model the emails as directed interactions over time using a Cox multiplicative intensity model and covariates that incorporate the past emailing behavior. The main results show homophily: the covariates gender, seniority and department are the most predictive effects of email interactions, together with a sender's own previous behavior and who emailed him recently.

Robinson and Priebe [125] analyze a version of the Enron dataset where each email got a topic assigned while the node-related covariates are omitted. The authors use a dynamic random dot product model for networks with attributed edges to identify emerging and disappearing

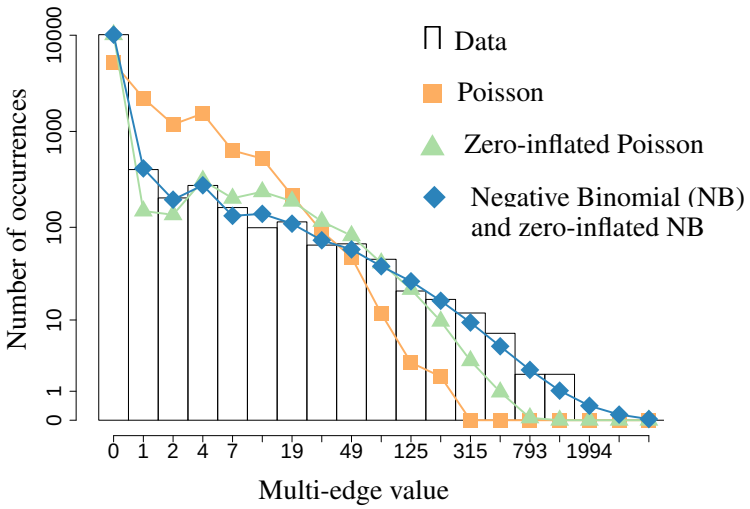


Figure 5.2: Observed versus expected email counts for maximum-likelihood fits of four different models satisfying Definition 2.

Model for the multi-edges A_{ij}	Degrees of freedom	Residual deviance	Relative change
1-parameter Poisson	1	232087	-
Poisson	153	142031	-39%
Zero-inflated Poisson	154	57070	-37%
Negative Binomial (NB)	154	12671	-19%
Zero-inflated NB	155	12671	0%

Table 5.3: Goodness-of-fit versus model complexity for the models in Figure 5.2 (residual deviance relative to a saturated Negative Binomial model with $r \to \infty$).

communities. The authors detect change points in the emailing behavior that match major events related to the collapse of the company.

In contrast to our work, Perry and Wolfe [120] and Robinson and Priebe [125] incorporate the time-dependence of the edges into the inference procedure and hence, their results focus on the time-dependent fluctuations of the emailing behavior. In agreement with our findings, Perry and Wolfe [120] identify the covariates gender, seniority and department as informative for the email interactions.

5.3.2 Elicitation of the model and deriving the p-values

To model this network, we will use the full flexibility afforded by Definition 2. Following the four steps described in Section 5.1, we determine a p-value corresponding to each covariate.

Step 1: To construct a suitable model for the observed multi-edges A_{ij}, we compare four different distributions satisfying the assumptions of Definition 2: $\mathrm{Poisson}(\pi_i\pi_j)$, $\mathrm{NegativeBinomial}(\pi_i\pi_j, r)$ with common shape parameter r, and zero-inflated versions of both.

Figure 5.2 shows how well the four distributions model the observed multi-edges. For each distribution, we fit a generalized linear model by maximum likelihood estimation. In Figure 5.2, we contrast the expected number of occurrences that two employees have exchanged $0, 1, 2, \ldots$ emails (denoted by the multi-edge value) under each model, with the observed number of occurrences. Even without zero-inflation, the Negative Binomial distribution yields a good fit, particularly in the right tail.

A formal model comparison via suitable likelihood ratio tests [26] confirms this: the Negative Binomial distribution achieves the best balance between fitting the observed data and model complexity. In Table 5.3, we contrast the model complexity (degrees of freedom) with the fit of the models to the observed data using the residual deviance D:

$$D = 2(l(\boldsymbol{A}; \boldsymbol{A}) - l(\mathbb{E}\,\boldsymbol{A}; \boldsymbol{A})),$$

with l being the log-likelihood. As the saturated model throughout, we use the Poisson log-likelihood $l(\mathbb{E}\,\boldsymbol{A}; \boldsymbol{A})$ with $\mathbb{E}\,\boldsymbol{A} = \boldsymbol{A}$; even when comparing with a Negative Binomial model since it holds that for all i, j

$$\lim_{r\to\infty} \mathrm{NegativeBinomial}(\mathbb{E}\,A_{ij}, r) = \mathrm{Poisson}(\mathbb{E}\,A_{ij}).$$

When comparing the Poisson and the Negative Binomial model each with the respective zero-inflated versions, we compare nested models and may therefore apply a classical likelihood ratio test. In [26], the authors explain that a Negative Binomial model is a compound Poisson distribution and show that

$$D(\mathrm{NegativeBinomial}) - D(\mathrm{Poisson}) \xrightarrow{d} \chi^2_1;$$

allowing us to apply a classical likelihood ratio test when comparing the Poisson with the Negative Binomial model. For a list of all likelihood functions see Appendix D.1. As a consequence

of all pairwise comparisons, we choose the model

$$A_{ij} \sim \text{NegativeBinomial}(\pi_i \pi_j, r). \tag{5.2}$$

Step 2: To verify the assumptions of Definition 2 for our data, we first assess Assumptions 1 and 2 exactly as before. Computing quartiles Q_1–Q_3 of the degrees—$68, 200, 564$—we see that Q_3/Q_2 and $Q_1/\sqrt{Q_2}$ are both of order one. Assumption 3 ($\max_i \pi_i/\sqrt{n}$ shrinking) can be analogously assessed via $Q_3/(n\sqrt{Q_2})$. Assumptions 4 and 5 require $\text{Var}\, A_{ij}/\,\mathbb{E}\, A_{ij} = 1 + \pi_i\pi_j/r$ and $\mathbb{E}\Big[(A_{ij} - \mathbb{E}\, A_{ij})^3\Big]\Big/\,\text{Var}\, A_{ij} = 1 + 2\pi_i\pi_j/r$ to be bounded. To assess this, we observe that a maximum-likelihood estimate of r [26] yields $\hat{r} = 0.047$, while the first three quartiles of $\widehat{\mathbb{E}\, A_{ij}}$ are respectively $0.16, 0.59, 2.1$. The ratio of the number of communities K over n is below 0.02 for all covariates, but first name initial with $K/n = 0.1111$ (see Table 5.4 for values of K).

Step 3: To estimate b and s in Theorem 4.2.2, we substitute $\hat{\pi}_i$ for π_i in Eqs. (4.4) and (4.24) exactly as before. Recall, however, that to estimate s we also require an estimate of $\text{Var}\, A_{ij}$ in Eq. (4.24). Under the parametrization of Eq. (5.2), it follows that

$$\text{Var}\, A_{ij} = \pi_i\pi_j(1 + \pi_i\pi_j/r). \tag{5.3}$$

Thus, $\text{Var}\, A_{ij}$ can be estimated by substituting $\hat{\pi}_i$ for π_i and $\hat{r}$ for r in (5.3). This yields the required estimators $\hat{b}$ and $\hat{s}$.

Step 4: To calculate p-values, we must first compute $(\widehat{Q} - \hat{b})/\hat{s}$ for each covariate. In advance of our analysis, we would expect that employee gender, seniority, and department might reflect aspects of community structure in email interactions. In contrast, we would expect covariates based on the first or last name of each individual to be non-informative. Figure 5.1 illustrates, in decreasing order of $(\widehat{Q} - \hat{b})/\hat{s}$, the observed structure of our data when grouped by covariate.

5.3.3 Results

Table 5.4 reports two approximate p-values per covariate, in contrast to the previous section. The first of these derives (via Eq. (5.1)) from Theorem 4.2.2, which shows the limiting distribution of $(\widehat{Q} - \hat{b})/\hat{s}$ under the assumed model to be a standard Normal. The second is based on 10^7 replicates of the parametric bootstrap, whereby we fit a negative Binomial model to the data, simulate from the fitted model and obtain the probability of the observed value of modularity under the empirical finite-sample distribution. Table 5.4 indicates that our asymptotic theory is conservative in this setting, leading as it does here to larger p-values than the bootstrap.

Covariate (no. groups)	$(\widehat{Q} - \hat{b})/\hat{s}$	p-value	
		Eq. (5.1)	Bootstrap
Department (3)	6.17	$< 10^{-6}$	$< 10^{-6}$
Seniority (3)	3.14	9×10^{-4}	8×10^{-6}
Gender (2)	2.36	9×10^{-3}	2×10^{-3}
First name initial (17)	0.74	2×10^{-1}	2×10^{-1}
Last name initial (3)	-0.46	7×10^{-1}	7×10^{-1}

Table 5.4: Analysis of the data of Figure 5.1, using modularity derived from multiple covariate-based community assignments.

Finally, considering these p-values in more detail, we see from Table 5.4 that for the covariates of department, gender, and seniority, all p-values fall below 1% (leading to a corrected total of 5% after adjusting for multiple comparisons). In contrast, we obtain large p-values for first- and last-name covariates (p-value $\geq 20\%$). This matches our expectations that department, gender, and seniority are likely to have an impact on email interactions, while there is no obvious reason why this should hold for name-related covariates.

5.4 Discussion

In this chapter, we demonstrated how to turn the theory derived in Chapters 3 and 4 into a methodology useful in practice. We here analyzed binary networks and a network with multiple edges to illustrate the flexibility of our method. The data analysis is conducted in two steps: validation and showcase.

First, we validated our methodology using four simple networks that are often used as benchmark examples for community detection [1, 59, 104, 108]. In this context, we discussed the behavior of modularity and the corresponding p-values under the null model of Definition 2 that does not support any community structure, via a parametric bootstrap. In addition, we analyzed the behavior of modularity and the corresponding p-values when the community assignments are informative: we defined them based on the observed covariates that have previously been used by others as ground truth for community detection [1, 59, 104, 108]. As a result, we observed that the data analysis on the benchmark networks aligns with the results in Theorem 4.2.2.

Second, we evaluated communities in a multi-edge email network. Our methodology identified employee seniority, gender, and company department as informative community assignments; in agreement with previous work by Perry and Wolfe [120]. In contrast, when running our method for the name related covariates we obtained large p-values (p-value ≥ 0.2). In this context, we applied a parametric bootstrap to obtain empirical p-values for each of the covariates as a second validation step, illustrating that our p-values are conservative.

Chapter 6

Summary, discussion, and future work

Here we provide a summary of the contributions of the thesis (Chapters 3–5). We then present a brief discussion of this and other prominent challenges in network modeling, along with possible avenues for future work (thereby extending Chapter 2).

6.1 Summary of our contributions

In this thesis, we contributed to the field of statistical inference for networks by deriving a new method to improve our understanding of community structure for large networks. We developed a general framework that enables us to model networks with node-specific differences but a lack of community structure. In this framework, we analyzed the asymptotic behavior of modularity—a well-known quality measure for community structure—and thereby derived a theoretical foundation for modularity for the first time. As a consequence, we obtained a methodology to identify which covariates serve as informative community structures, and then turned these theoretical results into a practical method.

For the framework, we extracted from the degree-based model [30] all properties essential to model node-specific differences with a lack of community structure. This led to a generalization of the degree-based model for simple networks to a nonparametric family of models, covering weighted, multi-edge, and power-law networks. Fitting these models, we generalized estimators discussed in [119] for this more general setting and derived limit theorems describing their asymptotic properties. Some of the later results extended work by [116].

To build the theoretical foundation of modularity, we first derived a statistical interpretation of the function itself. We showed that modularity is an estimator of a population quantity that contrasts the observed with the expected edges under the assumption of a degree-based model. As a consequence, we derived its asymptotic properties as an estimator, showing that

when appropriately standardized, it converges in distribution to a $\mathrm{Normal}(0, 1)$ random variable. Using the standardized modularity as a test statistic, we assessed whether an observed community structure leads to a modularity value that is unlikely under a degree-based model of no community structure. Due to the convergence in distribution result, we could quantify the Type I error that we falsely identify an observed community structure as informative.

We concluded the thesis with a data analysis where we demonstrated how to turn our theoretical results into a practical method. We first evaluated the method using four benchmark networks that have previously been used as ground truth for community detection. Using parametric bootstrap, we demonstrated our theoretical result that under a degree-based model, modularity is approximately $\mathrm{Normal}(0, 1)$ distributed for large networks. We then analyzed the behavior of modularity for covariates that have been reported by others to be informative community structures, illustrating that our method leads to small p-values in this case. In the second section of the data analysis, we evaluated communities in a multi-edge email network, demonstrating that our method may identify informative from non-informative covariates.

6.2 Discussion and future work

6.2.1 Community structure in networks

Networks have richer and more varied structure than can be described by a single "best" community assignment. To reflect this, we have introduced in this thesis an approach which exploits the structural information captured by covariates, each of which may describe different aspects of community structure in the data. In contrast to community *detection* per se, this approach allows us to assess the significance of a given, interpretable community assignment with respect to the observed network structure. As we have demonstrated in the data analysis examples, our method leads to the identification of structurally significant community assignments, ultimately yielding a better understanding of the network under study.

To advance the state of the art in network analysis, we as a research community must use the explanatory power of signed residual statistics as modularity to understand the effects of multiple observed communities on network structure. Our work here represents a first step in this direction: we have used the explanatory power of modularity to assess the significance of observed community structure relative to a null model. This opens the door to more advanced uses of multiple observed community assignments within formal statistical modeling frameworks. This is an important next step, since we have seen clear evidence here that multiple

groupings may explain different aspects of a network's community structure.

6.2.2 Network models with higher dimensionality

Due to technological advances we are able to collect data that are increasingly large and diverse in structure. To fully exploit these rich data, there is a strong need for network models to catch up in their dimensionality, and for us to derive the asymptotic properties of these models such that we can deliver statistical guarantees. In contrast to classical statistics, the observations may neither be identical nor independent; and there is no natural ordering inherited in the data and no means of geometry, as is the case for time series or spatial statistics. As a result, a key challenge here is to introduce high-dimensional models that reflect the unique structure inherited in networks.

As a start, we would suggest to introduce a dynamic network model with a memory. Most dynamic network models assume the networks at different time points to be conditionally independent given the parameters or latent positions [84, 47, 131, 147, 149]. There are few exceptions; e.g., the dynamic model by Snijders [134] and the temporal ERGM by Hanneke et al. [64], which suffer from high computational cost or degeneracy. For many applications though it is unrealistic to assume that generating a new connection is as likely as keeping a given connection, conditioned on the same parameters or latent positions.

6.2.3 Quantifying goodness-of-fit of network models

A fundamental question in modern science is how to draw conclusions from complex and high-dimensional data [55, 65]. While networks enable us to model complex dependencies between entities, the lack of goodness-of-fit tests for network models makes the results unreliable. Most current approaches for goodness-of-fit of networks either rely on summery statistics [71], or use classical methods without extending the theoretical foundation to the high-dimensional network setting [67, 71]. The classical goodness-of-fit statistics suffer from the fact that the asymptotic properties of classical methods designed for a fixed number of parameters, may be quite different when the number of parameters increases with the number of observations [44], as is the case in many network models.

For a statistician, it is natural to generalize the classical likelihood ratio test to a network setting. Most network models can be written as a likelihood function. To then test their goodness-of-fit using a likelihood ratio test statistic is intuitive and has proven to be robust in other settings. However, we need to derive the asymptotic distribution of the likelihood ratio test

statistic under the high-dimensional network models. Simons and Yao show asymptotic normality for the linear combination of model parameters where the number of parameters grows in the sample size [133]. The open problem is to generalize this result to serve for goodness-of-fit in a network setting.

Appendix A

Mathematical preliminaries

A.1 Probabilistic order notation

To discuss the properties of the sequence of random networks, we introduce standard notions of convergence in probability. For the non-random properties, we introduce the order notation; starting in Definition 3 with a notation for one sequence to be asymptotically negligible compared to another sequence.

Definition 3 (o notation)**.** *Let $\{a_n\}_{n\in N}$ and $\{b_n\}_{n\in N}$ be two sequences of real numbers. We write*

$$a_n = o(b_n) \quad \textit{if} \quad \frac{a_n}{b_n} \to 0 \quad \textit{as } n \to \infty.$$

Definition 4 (ω notation)**.** *Let $\{a_n\}_{n\in N}$ and $\{b_n\}_{n\in N}$ be two sequences of real numbers. We write*

$$a_n = \omega(b_n) \quad \textit{if} \quad \frac{a_n}{b_n} \to \infty \quad \textit{as } n \to \infty.$$

We now define below a notation for a sequence a_n to be at most asymptotically equivalent to another sequence b_n.

Definition 5 ($\mathcal{O}$ notation)**.** *Let $\{a_n\}_{n\in N}$ and $\{b_n\}_{n\in N}$ be two sequences of real numbers. We write*

$$a_n = \mathcal{O}(b_n) \quad \textit{if there exist} \quad M, N > 0 \textit{ such that } \left|\frac{a_n}{b_n}\right| < M \textit{ for all } n > N.$$

Note that $a_n = o(b_n)$ implies $a_n = \mathcal{O}(b_n)$ but not vise versa.

Definition 6 (Ω notation)**.** *Let $\{a_n\}_{n\in N}$ and $\{b_n\}_{n\in N}$ be two sequences of real numbers. We write*

$$a_n = \Omega(b_n) \quad \textit{if there exist} \quad M, N > 0 \textit{ such that } \left|\frac{a_n}{b_n}\right| > M \textit{ for all } n > N.$$

Definition 7 (Θ notation)**.** *Let $\{a_n\}_{n\in N}$ and $\{b_n\}_{n\in N}$ be two sequences of real numbers. We write*

$$a_n = \Theta(b_n) \quad \textit{if there exist} \quad M_1, M_2, N > 0 \textit{ such that } M_1 > \left|\frac{a_n}{b_n}\right| > M_2 \textit{ for all } n > N.$$

Definition 8 (Convergence in probability)**.** *Let X and $\{X_n\}_{n\in N}$ be a random variable and a sequence of random variables, respectively. Let both be defined on the same probability space. We say that X_n converges in probability to X, written $X_n \xrightarrow{P} X$ if for any $\epsilon > 0$, it holds that*

$$P(|X_n - X| < \epsilon) \to 1 \quad \textit{as } n \to \infty.$$

We say that an event E_n holds *with high probability*, if $P(E_n) \to 1$ as $n \to \infty$.

The notion of convergence in probability enables us to now introduce the order notation for random variables; which simplifies the proofs in the chapters below significantly.

Definition 9 (o_P notation)**.** *Let $\{X_n\}_{n\in N}$ and $\{Y_n\}_{n\in N}$ be two sequences of random variables. We write*

$$X_n = o_P(Y_n) \quad \textit{if} \quad \frac{X_n}{Y_n} \xrightarrow{P} 0 \quad \textit{as } n \to \infty.$$

Definition 10 ($\mathcal{O}_P$ notation)**.** *Let $\{X_n\}_{n\in N}$ and $\{Y_n\}_{n\in N}$ be two sequences of random variables. We write*

$$X_n = \mathcal{O}_P(Y_n)$$

if for every $\epsilon > 0$ there exist $M, N > 0$ such that

$$P\left(\left|\frac{X_n}{Y_n}\right| < M\right) > 1 - \epsilon \quad \textit{for all } n > N.$$

A.2 Standard results on convergence of random variables

A.2.1 Definitions

To address the random properties, we need to introduce a notion of the convergence of a sequence of random variables. While for a sequence of real numbers converging is uniquely defined, for random variables there are several notions; of which we have defined above convergence in probability, and define here convergence in distribution, in mean and almost sure convergence.

Definition 11 (Convergence in distribution)**.** *Let a random variable X and a sequence of random variables $\{X_n\}_{n\in N}$ be defined on a common probability space. We say that X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, if it holds that*

$$P(X_n \leq x) \to P(X \leq x) \quad \text{as } n \to \infty$$

for all x at which the cumulative distribution function of X is continuous.

The best-known result for convergence in distribution is the central limit theorem that says that the average of n *iid* random variables converges in distribution to a $\mathrm{Normal}(0, 1)$ random variable (given certain conditions).

Definition 12 (Almost sure convergence)**.** *Let X and $\{X_n\}_{n\in N}$ be a random variable and a sequence of random variables, respectively. Let both be defined on the same probability space. We say that X_n converges almost surely to X, written $X_n \xrightarrow{a.s.} X$ if it holds that*

$$P(\{X_n \to X\}) = 1.$$

Note that almost sure convergence implies convergence in probability and convergence in distribution: $X_n \xrightarrow{a.s.} X \Rightarrow X_n \xrightarrow{P} X, X_n \xrightarrow{d} X$ while the reverse implication does not hold. Almost sure convergence is therefore often coined strong convergence.

Definition 13 (Convergence in mean)**.** *Let X and $\{X_n\}_{n\in N}$ be a random variable and a sequence of random variables, respectively. We say that X_n converges in mean to X, written $X_n \xrightarrow{E} X$ if for any $\epsilon > 0$, it holds that*

$$\mathbb{E}(|X_n - X|) \to 0 \quad \text{as } n \to \infty.$$

A.2.2 Standard results

The notions of the convergence of random variables are strongly related, as the following theorem illustrates.

Theorem A.2.1 (Relation between the convergence notions [39, p. 3, 13, 88])**.** *Let X and $\{X_n\}_{n\in N}$ be a random variable and a sequence of random variables, respectively. Then,*

- *If $X_n \xrightarrow{d} X$, then $X_n = \mathcal{O}_P(1)$.*
- *If $X_n \xrightarrow{E} X$, then $X_n \xrightarrow{P} X$.*
- *If $X_n \xrightarrow{P} X$ and $\{X_n\}_{n\in N}$ are uniform integrable, then $X_n \xrightarrow{E} X$.*

Furthermore, the different types of convergence for random variable can be discussed together as the next theorem illustrates.

Theorem A.2.2 (Slutsky's theorem [39, p. 4])**.** *Let X and Y, and $\{X_n\}_{n\in N}$ and $\{Y_n\}_{n\in N}$ be a random variable and a sequence of random variables, respectively; and let $c \in \mathbb{R}$. Then,*

- *If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} c$, then $X_n \cdot Y_n \xrightarrow{d} c \cdot X$.*
- *If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} c \neq 0$, then $\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$.*
- *If $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{P} c$, then $X_n + Y_n \xrightarrow{d} X + c$.*

Theorem A.2.3 (Cramér–Wold theorem [39, p.9])**.** *Let X and $\{X_n\}_{n\in N}$ be a r-dimensional random vector and a sequence of r-dimensional random vectors,respectively. Then,*

$$X_n \xrightarrow{d} X \quad \text{if and only if} \quad \forall c \in \mathbb{R}^r : c'X_n \xrightarrow{d} c'X.$$

Theorem A.2.4 (Taylor expansion in probability [25, p. 201])**.** *Let us assume that X and Y are random variables, $f(X,Y)$ is a mapping and the following two assumptions hold. First, $\|(X,Y) - (h_X, h_Y)\|_2 = \mathcal{O}_p(r_n)$ where r_n goes to 0 as n goes to ∞ and second, the partial derivatives $\frac{\partial f}{\partial X_i}$ are continuous in a neighborhood of $\boldsymbol{h} = (h_X, h_Y)^T$. Then, we obtain the Taylor expansion in probability for a mapping $f(X,Y)$ at $\boldsymbol{h}$:*

$$\begin{aligned} f(X,Y) &= f(\boldsymbol{h}) + \partial_X f(\boldsymbol{h})(X - h_X) + \partial_Y f(\boldsymbol{h})(Y - h_Y) \\ &+ \frac{1}{2}\Big[\partial^2_{XX} f(\boldsymbol{h})(X - h_X)^2 + \partial^2_{YY} f(\boldsymbol{h})(Y - h_Y)^2 \\ &+ 2\partial^2_{XY} f(\boldsymbol{h})(X - h_X)(Y - h_Y)\Big] + o_p\big(r_n^2\big). \end{aligned}$$

Theorem A.2.5 (Lindeberg-Feller Central Limit Theorem [17, p. 359ff])**.** *Let $X_1, X_2, \ldots$ be independent random variables with $\operatorname{Var} X_i < \infty$ for all i. Under the Lindeberg condition that*

$$\forall \epsilon > 0 : \frac{1}{\sum_{i=1}^n \operatorname{Var} X_i} \sum_{i=1}^n \mathbb{E}\Big((X_i - \mathbb{E}\, X_i)^2 \delta_{|X_i - \mathbb{E}\, X_i| > \epsilon s_n}\Big) \xrightarrow{n} 0,$$

it follows that

$$\frac{1}{\sqrt{\sum_{i=1}^n \operatorname{Var} X_i}} \sum_{i=1}^n (X_i - \mathbb{E}\, X_i) \xrightarrow{d} \operatorname{Normal}(0,1).$$

Theorem A.2.6 (Lyapunov condition implies the Lindeberg condition [17, p. 362])**.** *Let $X_1, X_2, \ldots$ be independent random variables with $\operatorname{Var} X_i < \infty$ for all i. From the Lyapunov condition which for $s_n^2 = \sum_{i=1}^n \operatorname{Var} X_i$ states that*

$$\exists \delta > 0 : \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\Big(|X_i - \mathbb{E}\, X_i|^{2+\delta}\Big) \xrightarrow{n} 0;$$

implies the Lindeberg condition.

Appendix B

Supporting material for Chapter 3

B.1 Lemmas for proofs in Chapter 3

B.1.1 Lemmas for proof of Theorem 3.2.1

Lemma B.1.1. *Consider Assumptions 1 and 4. Then, it holds that*

$$\frac{\mathbb{E}\, d_i - \pi_i \sqrt{\mathbb{E}\, \|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} = \mathcal{O}\bigg(\frac{\max_j \pi_j - \pi_i}{\sqrt{n}}\bigg).$$

Proof. Since we know from Eqs. (3.2) and (3.3) that $\mathbb{E}\, d_i = \pi_i(\|\boldsymbol{\pi}\|_1 - \pi_i)$ and $\mathbb{E}\, \|\boldsymbol{d}\|_1 = \|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2$, we may write

$$\frac{\mathbb{E}\, d_i - \pi_i \sqrt{\mathbb{E}\, \|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} = \frac{\pi_i \|\boldsymbol{\pi}\|_1 \Big[1 - \sqrt{1 - \|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2}\Big] - \pi_i^2}{\sqrt{\operatorname{Var} d_i}}. \tag{B.1}$$

To apply a Taylor expansion of $\sqrt{1-x}$ for $x = \|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2$, we first need to show that x converges to 0. Considering $\tilde{\boldsymbol{\pi}} = \boldsymbol{\pi} / \max_j \pi_j$, we can conclude from $\|\tilde{\boldsymbol{\pi}}\|_2^2 \leq \|\tilde{\boldsymbol{\pi}}\|_1$ that

$$\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} = \frac{(\max_j \pi_j)^2 \|\tilde{\boldsymbol{\pi}}\|_2^2}{(\max_j \pi_j)^2 \|\tilde{\boldsymbol{\pi}}\|_1^2} \leq \frac{1}{\|\tilde{\boldsymbol{\pi}}\|_1} = \frac{\max_j \pi_j}{\|\boldsymbol{\pi}\|_1}. \tag{B.2}$$

Assumption 1 implies that $\max_j \pi_j / \|\boldsymbol{\pi}\|_1 = \mathcal{O}(1/n)$, and thus we conclude

$$\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} = \mathcal{O}\bigg(\frac{1}{n}\bigg). \tag{B.3}$$

This allows us to apply a convergent Taylor expansion of $\sqrt{1-x}$ at 0 in Eq. (B.1):

$$\begin{aligned}
&\frac{\mathbb{E}\, d_i - \pi_i\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}{\sqrt{\operatorname{Var} d_i}} \\
&= \frac{\pi_i\|\boldsymbol{\pi}\|_1\Big[1-\Big(1-\|\boldsymbol{\pi}\|_2^2/2\|\boldsymbol{\pi}\|_1^2+o\Big(\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2\Big)\Big)\Big]-\pi_i^2}{\sqrt{\operatorname{Var} d_i}} \\
&= \frac{\pi_i\Big[\|\boldsymbol{\pi}\|_2^2/2\|\boldsymbol{\pi}\|_1+o\Big(\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1\Big)\Big]-\pi_i^2}{\sqrt{\operatorname{Var} d_i}} \\
&\le \frac{\pi_i[\max_j \pi_j/2+o(\max_j \pi_j)]-\pi_i^2}{\sqrt{\operatorname{Var} d_i}} \qquad \text{(see Eq. (B.2))} \\
&= \Theta\left(\frac{\pi_i(\max_j \pi_j-\pi_i)}{\sqrt{\mathbb{E}\, d_i}}\right) \qquad \text{(Assumption 4)} \\
&= \Theta\left(\frac{\sqrt{\pi_i}(\max_j \pi_j-\pi_i)}{\sqrt{\|\pi\|_1-\pi_i}}\right) \\
&= \mathcal{O}\left(\frac{\max_j \pi_j-\pi_i}{\sqrt{n}}\right). \qquad \text{(Assumption 1)}
\end{aligned}$$

□

Lemma B.1.2. *Consider Assumptions 2–5. Then,* $\sqrt{\|\boldsymbol{d}\|_1}-\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}=\mathcal{O}_P(1)$.

Proof. Observe that the square root function has one continuous derivative at 1. A Taylor expansion in probability (see Appendix A.2) of $\sqrt{\|\boldsymbol{d}\|_1/\,\mathbb{E}\,\|\boldsymbol{d}\|_1}$ about 1 requires in addition that

I. $\exists a\in\mathbb{R}:\|\boldsymbol{d}\|_1/\,\mathbb{E}\,\|\boldsymbol{d}\|_1=a+\mathcal{O}_P(r_n)$; with

II. $r_n\to 0$ as $n\to\infty$.

I. It follows from Chebyshev's inequality that

$$\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\,\|\boldsymbol{d}\|_1}=1+\mathcal{O}_P\left(\frac{\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}}{\mathbb{E}\,\|\boldsymbol{d}\|_1}\right). \tag{B.4}$$

II. As a consequence of I., $r_n=\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}/\,\mathbb{E}\,\|\boldsymbol{d}\|_1$. From Eq. (3.3) and Assumption 2 ($\Rightarrow \mathbb{E}\, d_i\to\infty$) it follows that $\mathbb{E}\|\boldsymbol{d}\|_1\to\infty$. Since A_{ij} are independent for $i<j$, and since we assume $\operatorname{Var} A_{ij}/\,\mathbb{E}\, A_{ij}=\Theta(1)$ (Assumption 4), it holds that

$$\begin{aligned}
\frac{\operatorname{Var}\|\boldsymbol{d}\|_1}{\mathbb{E}\,\|\boldsymbol{d}\|_1} &= \frac{\operatorname{Var}\Big(2\sum_{j=1}^n\sum_{i<j}A_{ij}\Big)}{\mathbb{E}\Big(2\sum_{j=1}^n\sum_{i<j}A_{ij}\Big)} \\
&= \frac{4\sum_{j=1}^n\sum_{i<j}\operatorname{Var}(A_{ij})}{2\sum_{j=1}^n\sum_{i<j}\mathbb{E}(A_{ij})} \\
&= \Theta(1).
\end{aligned} \tag{B.5}$$

It follows that the ratio $\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}/\mathbb{E}\|\boldsymbol{d}\|_1 \to 0$.

We now can apply a convergent Taylor expansion in probability:

$$\sqrt{\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1}} = 1 + \frac{1}{2}\left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1\right) + o_P\left(\frac{\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}}{\mathbb{E}\|\boldsymbol{d}\|_1}\right)$$

$$\Leftrightarrow \quad \sqrt{\|\boldsymbol{d}\|_1} - \sqrt{\mathbb{E}\|\boldsymbol{d}\|_1} = \frac{\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}\left[\frac{1}{2}\left(\frac{\|\boldsymbol{d}\|_1 - \mathbb{E}\|\boldsymbol{d}\|_1}{\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}}\right) + o_P(1)\right]. \tag{B.6}$$

Since the term $\|\boldsymbol{d}\|_1/2 = \sum_{j=1}^n \sum_{i<j} A_{ij}$ is a sum of independent random variables, we apply the Lindeberg–Feller central limit theorem (see Appendix A.2) analogously to Term T_1: From Assumptions 2–5, it follows that

$$\frac{\|\boldsymbol{d}\|_1 - \mathbb{E}\|\boldsymbol{d}\|_1}{\sqrt{\operatorname{Var}\|\boldsymbol{d}\|_1}} \xrightarrow{d} \operatorname{Normal}(0,1).$$

Since $\operatorname{Var}\|\boldsymbol{d}\|_1/\mathbb{E}\|\boldsymbol{d}\|_1 = \Theta(1)$ by Eq. (B.5), we conclude from Eq. (B.6) the result of Lemma B.1.2; i.e., $\sqrt{\|\boldsymbol{d}\|_1} - \sqrt{\mathbb{E}\|\boldsymbol{d}\|_1} = \mathcal{O}_P(1)$. □

B.1.2 Lemmas for proof of Theorem 3.2.2

Lemma B.1.3. *Consider Assumptions 1, 2 and 4. Then, it holds that*

$$\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2} = \mathcal{O}_P\left(\frac{1}{n}\right).$$

Proof. First, from Chebyshev's inequality, and from Assumption 4, we know that

$$\frac{\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_2^2} = 1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right) \quad \text{and} \quad \frac{\|\boldsymbol{d}\|_1^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2} = 1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1^2}}\right). \tag{B.7}$$

In return, it follows that

$$\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2} = \frac{\mathbb{E}\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right)\right]\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1^2}}\right)\right]^{-1}.$$

We may apply a convergent Taylor expansion of $f(x) = (1+x)^{-1}$ at 1, since $x = 1\Big/\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1^2} = o(1)$. It follows that

$$= \frac{\mathbb{E}\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right)\right]\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1^2}}\right)\right]$$

$$= \frac{\mathbb{E}\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right)\right]. \qquad \left(\text{since } \|\boldsymbol{d}\|_2^2 \le \|\boldsymbol{d}\|_1^2\right) \tag{B.8}$$

Via straightforward algebraic computations, we obtain

$$\begin{aligned}\mathbb{E}\|\boldsymbol{d}\|_2^2 &= \sum_i \sum_{j\neq i} \sum_{l\neq i} \mathbb{E}(A_{ij}A_{il}) \\ &= \sum_i \mathbb{E}\, d_i\, \mathbb{E}\, d_i \cdot (1+o(1)) \\ &= \|\boldsymbol{\pi}\|_1^2 \|\boldsymbol{\pi}\|_2^2 \cdot (1+o(1)), \qquad \text{(Assumption 1)}\end{aligned} \tag{B.9}$$

and

$$\begin{aligned}\mathbb{E}\|\boldsymbol{d}\|_1^2 &= \mathrm{Var}\|\boldsymbol{d}\|_1 + (\mathbb{E}\|\boldsymbol{d}\|_1)^2 \\ &= \Theta(\mathbb{E}\|\boldsymbol{d}\|_1) + (\mathbb{E}\|\boldsymbol{d}\|_1)^2 \quad \text{(Assumption 4)} \\ &= \Theta\Big[(\mathbb{E}\|\boldsymbol{d}\|_1)^2\Big]. \quad \text{(Assumption 2)}\end{aligned} \tag{B.10}$$

We know from Eq. (B.8) that

$$\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1^2} = \frac{\mathbb{E}\|\boldsymbol{d}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_2^2}}\right)\right].$$

Combining Eqs. (B.9) and (B.10) and applying Assumption 1, it then follows that

$$\begin{aligned}&= \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\left[1 + \mathcal{O}_P\left(\frac{1}{\|\boldsymbol{\pi}\|_1\|\boldsymbol{\pi}\|_2}\right)\right] \\ &= \mathcal{O}_P\left(\frac{1}{n}\right). \qquad \text{(see Eq. (B.3))}\end{aligned}$$

□

Lemma B.1.4. *Consider Assumptions 1, 2 and 4, and assume* $A_{ij} \sim \mathrm{Poisson}(\pi_i\pi_j)$. *Then, it holds for the plug-in estimator* $\widehat{\mathrm{Var}\, d_i}$ *where we exchange each* π_i *in* $\mathrm{Var}\, d_i$ *by* $\hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}$ *that*

$$\frac{\widehat{\mathrm{Var}\, d_i}}{\mathrm{Var}\, d_i} \xrightarrow{P} 1.$$

Proof. For Poisson-distributed edges, $\mathbb{E}\, A_{ij} = \mathrm{Var}\, A_{ij}$ for all i, j. Hence, we obtain

$$\begin{aligned}\frac{\widehat{\mathrm{Var}\, d_i}}{\mathrm{Var}\, d_i} &= \frac{\widehat{\mathbb{E}\, d_i}}{\mathbb{E}\, d_i} \\ &= \frac{\hat{\pi}_1\|\hat{\boldsymbol{\pi}}\|_1 - \hat{\pi}_i^2}{\mathbb{E}\, d_i} \\ &= \frac{\frac{d_i}{\sqrt{\|\boldsymbol{d}\|_1}}\frac{\|\boldsymbol{d}\|_1}{\sqrt{\|\boldsymbol{d}\|_1}} - \frac{d_i^2}{\|\boldsymbol{d}\|_1}}{\mathbb{E}\, d_i} \\ &= \frac{d_i}{\mathbb{E}\, d_i}\left[1 - \frac{d_i}{\|\boldsymbol{d}\|_1}\right]\end{aligned}$$

$$= \left[1 + \mathcal{O}_P\left(\sqrt{\frac{\operatorname{Var} d_i}{(\mathbb{E}\, d_i)^2}}\right)\right]\left[1 - \frac{d_i}{\|\boldsymbol{d}\|_1}\right] \quad \text{(Chebyshev's inequality)}$$

$$= \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right]\left[1 - \frac{d_i}{\|\boldsymbol{d}\|_1}\right]. \quad \text{(Assumption 4)} \tag{B.11}$$

Furthermore, from Assumptions 1 ($n\pi_i/\|\boldsymbol{\pi}\|_1 = \mathcal{O}(1)$), 2 ($\Rightarrow \mathbb{E}\, d_i \to \infty$), and 4 ($\operatorname{Var} A_{ij} = \Theta(\mathbb{E}\, A_{ij})$), it follows that $\frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{d_i}{\|\boldsymbol{d}\|_1} \xrightarrow{P} 1$, as we will now show.

We write

$$\frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{d_i}{\|\boldsymbol{d}\|_1} = \underbrace{\left(\frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{\mathbb{E}\, d_i}{\|\boldsymbol{\pi}\|_1^2}\right)}_{c_n}\underbrace{\left(\frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2}\right)^{-1}}_{E_n}\underbrace{\left(\frac{d_i}{\mathbb{E}\, d_i}\right)}_{F_n}. \tag{B.12}$$

By Chebyshev's inequality and from Assumptions 2 and 4, we know that

$$F_n = \frac{d_i}{\mathbb{E}\, d_i} = 1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right).$$

For E_n, we will first establish the equivalence

$$\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\,\|\boldsymbol{d}\|_1} = \frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2} = \frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2}\left[1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]^{-1}$$

$$\Leftrightarrow \quad \frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2} = \frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\,\|\boldsymbol{d}\|_1}\left[1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right].$$

By Eq. (B.3), we know that from Assumption 1 it follows that $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$. Furthermore, by Chebyshev's inequality and from Assumptions 2 and 4, $\|\boldsymbol{d}\|_1/\mathbb{E}\,\|\boldsymbol{d}\|_1 \xrightarrow{P} 1$. Thus, it follows that

$$E_n = \frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2} = \frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\,\|\boldsymbol{d}\|_1}\left[1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right] = 1 + \mathcal{O}_P\left(\frac{1}{\min\left(n, \sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}\right)}\right). \tag{B.13}$$

For the non-random sequence $\{c_n; n \in \mathbb{N}\}$ in Eq. (B.12) it holds that

$$c_n = \frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{\mathbb{E}\, d_i}{\|\boldsymbol{\pi}\|_1^2}$$

$$= \frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{\pi_i\|\boldsymbol{\pi}\|_1}{\|\boldsymbol{\pi}\|_1^2}\left[1 - \frac{\pi_i}{\|\boldsymbol{\pi}\|_1}\right]$$

$$= \left[1 + \mathcal{O}\left(\frac{1}{n}\right)\right]. \quad \text{(Assumption 1)}$$

The inverse of a random variable which converges in probability to a constant c must in turn converge to $1/c$, as long as $c \neq 0$ [85, Theorem 2.1.3]. Furthermore, the product of two random variables, converging in probability to a constant c and a constant d respectively, itself converges to the product of the constants cd [85, Theorem 2.1.3]. Thus, it follows that

$$\frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\frac{d_i}{\|\boldsymbol{d}\|_1} = c_n E_n^{-1} F_n = 1 + \mathcal{O}_P\left(\frac{1}{\min_i\left(\sqrt{\mathbb{E}\, d_i}, n\right)}\right). \tag{B.14}$$

$$\Leftrightarrow \quad \frac{d_i}{\|\boldsymbol{d}\|_1} = \mathcal{O}_P\left(\frac{1}{n}\right). \quad \text{(Assumption 1)}$$

Recall from Eq. (B.11) that

$$\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} = \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right]\left[1 - \frac{d_i}{\|\boldsymbol{d}\|_1}\right].$$

In turn, we obtain the required result; i.e.,

$$\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} = 1 + \mathcal{O}_P\left(\frac{1}{\min_i\left(\sqrt{\mathbb{E}\, d_i}, n\right)}\right).$$

From Assumption 2 ($\pi_i = \omega(1/\sqrt{n})$), it follows that $\min_i \mathbb{E}\, d_i$ diverges. Hence, we have shown the required result that $\operatorname{Var} d_i$ can be consistently estimated by its plug-in estimator $\widehat{\operatorname{Var} d_i}$. □

Lemma B.1.5. *Consider Assumptions 1, 2 and 4, and assume* $A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j)$. *Then, it holds for the plug-in estimator* $\widehat{\operatorname{Var} d_i}$ *where we exchange each* π_i *in* $\operatorname{Var} d_i$ *by* $\hat{\pi}_i = d_i/\sqrt{\|\boldsymbol{d}\|_1}$ *that*

$$\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} \xrightarrow{P} 1.$$

Proof. For Bernoulli-distributed edges, we obtain $\operatorname{Var} d_i = \mathbb{E}\, d_i - \pi_i^2\|\boldsymbol{\pi}\|_2^2 + \pi_i^4$ [116]. We write

$$\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} = \frac{\hat{\pi}_i\|\hat{\boldsymbol{\pi}}\|_1 - \hat{\pi}_i^2 - \hat{\pi}_i^2\|\hat{\boldsymbol{\pi}}\|_2^2 + \hat{\pi}_i^4}{\pi_i\|\boldsymbol{\pi}\|_1 - \pi_i^2 - \pi_i^2\|\boldsymbol{\pi}\|_2^2 + \pi_i^4}.$$

It can easily been seen that $\hat{\pi}_i\|\hat{\boldsymbol{\pi}}\|_1 = d_i$ and $\|\hat{\boldsymbol{\pi}}\|_2^2 = \|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1$. It follows that

$$\begin{aligned}
&= \frac{d_i - d_i^2/\|\boldsymbol{d}\|_1 - d_i^2\|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1^2 + d_i^4/\|\boldsymbol{d}\|_1^2}{\pi_i\|\boldsymbol{\pi}\|_1 - \pi_i^2 - \pi_i^2\|\boldsymbol{\pi}\|_2^2 + \pi_i^4} \\
&= \frac{d_i - d_i^2/\|\boldsymbol{d}\|_1 - d_i^2\|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1^2 + d_i^4/\|\boldsymbol{d}\|_1^2}{\pi_i\|\boldsymbol{\pi}\|_1 - \pi_i^2\|\boldsymbol{\pi}\|_2^2} \cdot [1 + o(1)] \qquad \text{(Assumption 1)} \\
&= \frac{d_i[1 - d_i/\|\boldsymbol{d}\|_1] - d_i^2\left[\|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1^2 + d_i^2/\|\boldsymbol{d}\|_1^2\right]}{\pi_i\|\boldsymbol{\pi}\|_1 - \pi_i^2\|\boldsymbol{\pi}\|_2^2} \cdot [1 + o(1)]. \qquad \text{(B.15)}
\end{aligned}$$

We have seen in Eq. (B.21) that Assumptions 2 and 4 imply that

$$\frac{d_i}{\sqrt{\|\boldsymbol{d}\|_1}} = \frac{\mathbb{E}\, d_i}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right].$$

It follows from identical arguments that

$$\frac{d_i}{\|\boldsymbol{d}\|_1} = \frac{\mathbb{E}\, d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1}\left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right]. \qquad \text{(B.16)}$$

From Assumption 1, we conclude that

$$\begin{aligned}\frac{\mathbb{E}\, d_i}{\mathbb{E}\,\|\boldsymbol{d}\|_1} &= \frac{\pi_i(1-\pi_i/\|\boldsymbol{\pi}\|_1)}{\|\boldsymbol{\pi}\|_1\big(1-\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2\big)} \\ &= \frac{\pi_i}{\|\boldsymbol{\pi}\|_1}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \qquad \text{(see Eq. (B.23))} \\ &= \mathcal{O}\left(\frac{1}{n}\right). \qquad \text{(Assumption 1)} \end{aligned} \tag{B.17}$$

Combining Eqs. (B.16) and (B.17), it follows that

$$\frac{d_i}{\|\boldsymbol{d}\|_1} = \mathcal{O}_P\left(\frac{1}{n}\right).$$

It follows in turn that in combination with Eq. (B.15), we obtain

$$\begin{aligned}\frac{\widehat{\operatorname{Var} d_i}}{\operatorname{Var} d_i} &= \frac{d_i - d_i^2\|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1^2}{\pi_i\|\boldsymbol{\pi}\|_1 - \pi_i^2\|\boldsymbol{\pi}\|_2^2}\cdot[1+o_P(1)] \\ &= \underbrace{\frac{d_i}{\pi_i\|\boldsymbol{\pi}\|_1}}_{R_n}\cdot\underbrace{\frac{1-d_i/\|\boldsymbol{d}\|_1\|\boldsymbol{d}\|_2^2/\|\boldsymbol{d}\|_1}{1-\pi_i/\|\boldsymbol{\pi}\|_1\,\|\boldsymbol{\pi}\|_2^2}}_{S_n}\cdot[1+o_P(1)]. \end{aligned} \tag{B.18}$$

Term R_n:

$$\begin{aligned} R_n &= \frac{d_i}{\pi_i\|\boldsymbol{\pi}\|_1} \\ &= \frac{\mathbb{E}\, d_i}{\pi_i\|\boldsymbol{\pi}\|_1}\left[1+\mathcal{O}_P\left(\sqrt{\frac{\operatorname{Var} d_i}{(\mathbb{E}\, d_i)^2}}\right)\right] \qquad \text{(Chebyshev's inequality)} \\ &= \frac{\mathbb{E}\, d_i}{\pi_i\|\boldsymbol{\pi}\|_1}\left[1+\mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right] \qquad \text{(Assumption 4)} \\ &= 1+\mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right) \qquad \text{(Assumption 1)} \\ &= 1+o_P(1). \qquad \text{(Assumption 2)} \end{aligned}$$

Term S_n: We show the convergence of S_n from Eq. (B.18) in two steps:

1. $\frac{\|\boldsymbol{\pi}\|_1}{\pi_i}\,\frac{d_i}{\|\boldsymbol{d}\|_1} \xrightarrow{P} 1$;
2. $\left(\|\boldsymbol{\pi}\|_2^2\right)^{-1}\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1} \quad \xrightarrow{P} \quad 1.$

Step 1: This step follows analogously to Eq. (B.14) for $A_{ij} \sim \text{Poisson}(\pi_i\pi_j)$.

Step 2: We write the ratio of interest as

$$\left(\|\boldsymbol{\pi}\|_2^2\right)^{-1}\frac{\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{d}\|_1} = \left(\frac{\|\boldsymbol{d}\|_1}{\|\boldsymbol{\pi}\|_1^2}\right)^{-1}\cdot\left(\frac{\|\boldsymbol{d}\|_2^2}{\mathbb{E}\,\|\boldsymbol{d}\|_2^2}\right)\cdot\left(\frac{\mathbb{E}\,\|\boldsymbol{d}\|_2^2}{\|\boldsymbol{\pi}\|_2^2\|\boldsymbol{\pi}\|_1^2}\right) = L_n^{-1}M_n t_n.$$

Now, we analyze L_n, M_n and t_n in consecutive order. Under Assumptions 1, 2 and 4, we know that $L_n = \|\boldsymbol{d}\|_1/\|\boldsymbol{\pi}\|_1^2 \xrightarrow{P} 1$ (see Eq. (B.13)). Furthermore, combining Eqs. (B.7) and (B.9) enables us to conclude that $M_n = \|\boldsymbol{d}\|_2^2/\mathbb{E}\,\|\boldsymbol{d}\|_2^2 \xrightarrow{P} 1$ (under Assumptions 1 and 4). From Eq. (B.9), we know that under Assumption 1, the sequence $\{t_n; n \in \mathbb{N}\}$ converges to 1.

The inverse of a random variable which converges in probability to a constant c, must in turn converge to $1/c$, as long as $c \neq 0$ [85, Theorem 2.1.3]. Furthermore, the product of two random variables, converging in probability to a constant c and a constant d respectively, itself converges to the product of the constants cd [85, Theorem 2.1.3]. Thus, Step 2 follows.

Returning now to Eq. (B.18) and following the same argument, we conclude that $S_n \xrightarrow{P} 1$ and in turn, $\widehat{\operatorname{Var} d_i}/\operatorname{Var} d_i = R_n S_n[1 + o_P(1)] \xrightarrow{P} 1$ for Bernoulli-distributed edges (i.e., $A_{ij} \sim \text{Bernoulli}(\pi_i\pi_j)$). □

B.1.3 Lemma for proof of Corollary 3.2.1

Lemma B.1.6. *Consider Assumptions 1–5. Then, it holds that*

1. $\boldsymbol{D_{11}} \xrightarrow{n} \boldsymbol{I_r}$,
2. $\boldsymbol{m_{12}} \xrightarrow{d} \text{Normal}(\boldsymbol{0}, \boldsymbol{I_r})$,
3. $\boldsymbol{m_{13}} \xrightarrow{P} \boldsymbol{0}$.

Proof. Step 1: For the term $\boldsymbol{D_{11}}$, it holds for all i that

$$\sqrt{\frac{\operatorname{Var}(\sum_{l=r+1}^{n} A_{li})}{\operatorname{Var} d_i}} = \sqrt{1 - \frac{\operatorname{Var}(\sum_{l=1}^{r} A_{li})}{\operatorname{Var}(\sum_{l=1}^{n} A_{li})}}.$$

Furthermore, from Assumption 4 ($\operatorname{Var} A_{ij} = \Theta(\mathbb{E}\, A_{ij})$) we conclude for all i that

$$\begin{aligned}\frac{\operatorname{Var}(\sum_{l=1}^{r} A_{li})}{\operatorname{Var}(\sum_{l=1}^{n} A_{li})} &= \Theta\left(\frac{\sum_{l=1}^{r} \mathbb{E}\, A_{li}}{\sum_{l=1}^{n} \mathbb{E}\, A_{li}}\right) \\ &= \Theta\left(\frac{\pi_i \sum_{l=1}^{r} \pi_l}{\pi_i \|\boldsymbol{\pi}\|_1}\right) \\ &= \Theta\left(\sum_{l=1}^{r} \frac{\pi_l}{\|\boldsymbol{\pi}\|_1}\right).\end{aligned}$$

It follows further from Assumption 1 that

$$\frac{\operatorname{Var}(\sum_{l=1}^{r} A_{li})}{\operatorname{Var}(\sum_{l=1}^{n} A_{li})} = \mathcal{O}\left(\frac{r}{n}\right) \to 0. \tag{B.19}$$

In turn, $\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{li})}/\sqrt{\operatorname{Var} d_i} \to 1$ for all i. Hence, the diagonal matrix $\boldsymbol{D_{11}}$ converges to the identity matrix $\boldsymbol{I_r}$ in the operator norm.

Step 2: The term $\boldsymbol{m_{12}} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r})$, as we will now show by applying the Cramér–Wold theorem. The term $\boldsymbol{m_{12}}$ is a random vector depending on n, where each component is a sum of independent random variables. We will show now that, as a consequence, each component converges marginally in distribution to a $\mathrm{Normal}(0,1)$ random variable (by the same argument as in Theorem 3.2.1 for Term T_1). From Assumption 2 ($\Rightarrow \mathbb{E}\, d_i \to \infty$) and Assumption 4 ($\operatorname{Var} A_{ij} / \mathbb{E}\, A_{ij} = \Theta(1)$), it follows that $\operatorname{Var} d_i \to \infty$. Since in addition we assume the skewness of each edge A_{ij} to be bounded asymptotically (Assumption 5), the Lyapunov condition (for $\delta = 1$) is satisfied for each component. Hence, the Lindeberg–Feller central limit theorem (see Appendix A.2) lets us conclude that each component converges marginally in distribution to a $\mathrm{Normal}(0,1)$ random variable [17, p. 362].

Furthermore, the components of $\boldsymbol{m_{12}}$ are independent. It follows that for each $(c_1, \ldots, c_r) \in \mathbb{R}^r$ and $Y_u \overset{\text{iid}}{\sim} \mathrm{Normal}(0,1)$ for $u = 1, \ldots, r$, it holds that

$$\sum_{u=1}^{r} c_u \frac{\sum_{l=r+1}^{n} (A_{lu} - \mathbb{E}\, A_{lu})}{\sqrt{\operatorname{Var}(\sum_{l=r+1}^{n} A_{lu})}} \quad \xrightarrow{d} \quad \sum_{u=1}^{r} c_u Y_u.$$

Applying the Cramér–Wold theorem (see Appendix A.2), we conclude that $\boldsymbol{m_{12}} \xrightarrow{d} \mathrm{Normal}(\mathbf{0}, \boldsymbol{I_r})$.

Step 3: Finally, term $\boldsymbol{m_{13}} \xrightarrow{P} \mathbf{0}$, since by Chebyshev's inequality

$$\frac{\sum_{l=1}^{r} (A_{li} - \mathbb{E}\, A_{li})}{\sqrt{\operatorname{Var} d_i}} = \mathcal{O}_P\left(\sqrt{\frac{\operatorname{Var}(\sum_{l=1}^{r} A_{li})}{\operatorname{Var} d_i}} \right),$$

which in turn goes to 0 for all i, as seen in Eq. (B.19). □

B.1.4 Lemma for proof of Corollary 3.3.1

Lemma B.1.7. *Consider Assumptions 1, 2 and 4. Then,*

$$\hat{\pi}_i - \pi_i = \mathcal{O}_P\left(\frac{\pi_i}{\sqrt{\mathbb{E}\, d_i}} \right),$$
$$\frac{(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}{\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)} = \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}} \right).$$

Proof. First, we appeal to a Taylor expansion in probability of $\hat{\pi}_i = d_i / \sqrt{\|\boldsymbol{d}\|_1}$ (see Appendix A.2). Let $A = d_i / \mathbb{E}\, d_i$ and $B = (\|\boldsymbol{d}\|_1 - 2d_i) / \mathbb{E}(\|\boldsymbol{d}\|_1 - 2d_i)$. Observe that the function

$$\hat{\pi}_i = f(A, B) = \frac{\mathbb{E}\, d_i \; A}{\sqrt{2\, \mathbb{E}\, d_i \; A + \mathbb{E}(\|\boldsymbol{d}\|_1 - 2d_i) \; B}} \tag{B.20}$$

has continuous partial derivatives at $(1,1)'$. A Taylor expansion in probability of f requires in addition that $\sqrt{(A-1)^2+(B-1)^2} \xrightarrow{P} 0$. By Chebyshev's inequality, we know that

$$\begin{aligned}\sqrt{(A-1)^2+(B-1)^2} &= \sqrt{\left(\frac{d_i}{\mathbb{E}\, d_i}-1\right)^2+\left(\frac{\|\boldsymbol{d}\|_1-2d_i}{\mathbb{E}(\|\boldsymbol{d}\|_1-2d_i)}-1\right)^2}\\ &= \sqrt{\mathcal{O}_p\left[\operatorname{Var}\left(\frac{d_i}{\mathbb{E}\, d_i}\right)\right]+\mathcal{O}_p\left[\operatorname{Var}\left(\frac{\|\boldsymbol{d}\|_1-2d_i}{\mathbb{E}(\|\boldsymbol{d}\|_1-2d_i)}\right)\right]}\\ &= \sqrt{\mathcal{O}_p\left[\frac{\operatorname{Var} d_i}{(\mathbb{E}\, d_i)^2}\right]+\mathcal{O}_p\left[\frac{\operatorname{Var}(\|\boldsymbol{d}\|_1-2d_i)}{(\mathbb{E}(\|\boldsymbol{d}\|_1-2d_i))^2}\right]}.\end{aligned}$$

From Assumptions 2 and 4 ($\Rightarrow \mathbb{E}\, d_i \to \infty$, $\operatorname{Var} A_{ij}/\mathbb{E}\, A_{ij} = \Theta(1)$), it follows that $\sqrt{(A-1)^2+(B-1)^2} \xrightarrow{P} 0$.

We now can expand the function $f(A,B)$ in Eq. (B.20) in a convergent Taylor series around $(1,1)'$. In combination with Assumptions 2 and 4 we obtain

$$\frac{d_i}{\sqrt{\|\boldsymbol{d}\|_1}} = \frac{\mathbb{E}\, d_i}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}\left[1+\mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right]. \tag{B.21}$$

Furthermore, we conclude that

$$\begin{aligned}\frac{\mathbb{E}\, d_i}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}} &= \frac{\pi_i(1-\pi_i/\|\boldsymbol{\pi}\|_1)}{\sqrt{1-\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2}} && (B.22)\\ &= \pi_i\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]\left[1-\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]^{-1/2} && \text{(Assumption 1)}\\ &= \pi_i\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]\left[1+\mathcal{O}\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1}\right)\right] && \text{(Taylor expansion)}\\ &= \pi_i\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]. && \text{(see Eq. (B.3))} \quad (B.23)\end{aligned}$$

Combining Eqs. (B.21) and (B.23), it follows that

$$\hat{\pi}_i = \frac{d_i}{\sqrt{\|\boldsymbol{d}\|_1}} = \pi_i\left[1+\mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i}}\right)\right].$$

We conclude immediately the first result of Lemma B.1.7; i.e.,

$$\hat{\pi}_i - \pi_i = \mathcal{O}_P\left(\frac{\pi_i}{\sqrt{\mathbb{E}\, d_i}}\right). \tag{B.24}$$

In turn, the second result of Lemma B.1.7 follows; i.e.,

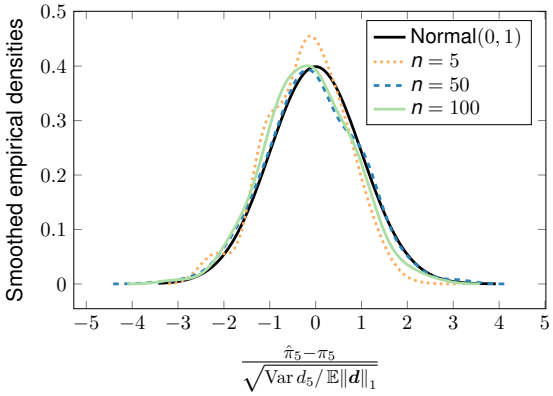
$$\begin{aligned}
&\frac{(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}{\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)} \\
&= \left[\frac{\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)}{(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}\right]^{-1} \\
&= \left[\frac{\pi_j}{\hat{\pi}_j - \pi_j} + \frac{\pi_i}{\hat{\pi}_i - \pi_i}\right]^{-1} \\
&= \left[\Omega\Big(\sqrt{\mathbb{E}\, d_i}\Big) + \Omega\Big(\sqrt{\mathbb{E}\, d_j}\Big)\right]^{-1} \qquad \text{(see Eq. (B.24))} \\
&= \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}}\right).
\end{aligned}$$

□

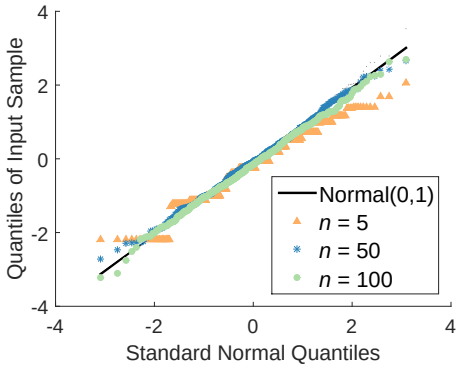
B.2 Simulations illustrating theorems in Chapter 3

B.2.1 Simulations illustrating Theorem 3.2.1

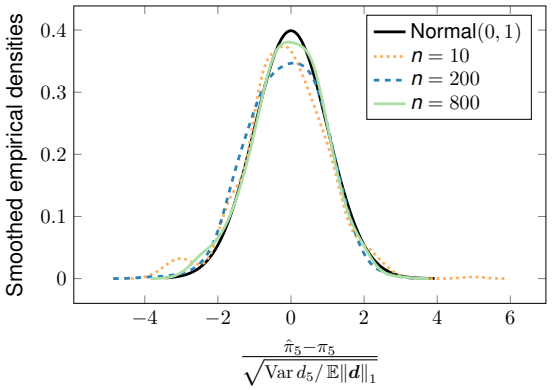
The following two simulations shall illustrate that the convergence in distribution is driven by the effective sample size rather than the number of nodes. We here display the results of repeating the simulation described in Section 3.4 for $\theta = 0.2$ with varied $\gamma = [0, 1)$ and node indices i. In both cases we observe that as n increases the difference between the smoothed empirical density of π_i and a $\mathrm{Normal}(0, 1)$ density shrinks. However, we see that the convergence here is faster (it takes fewer nodes to see the same convergence rate) than for $\gamma = 0.6$ since the network is less sparse. Figures B.1 and B.2 below illustrate this point with simulations from $\pi_i = i^{-0.2}$ and $\pi_i = 0.9\ i^{-0.2}$ for node 5 and 17, respectively.



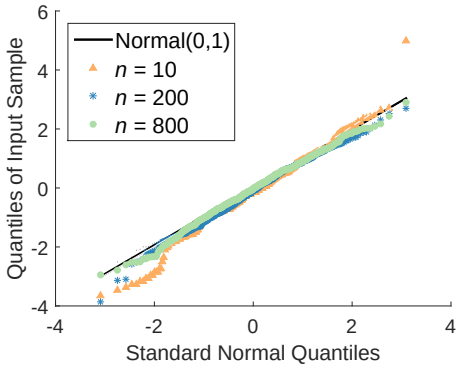
(a) Smoothed empirical densities



(b) Q-Q plot

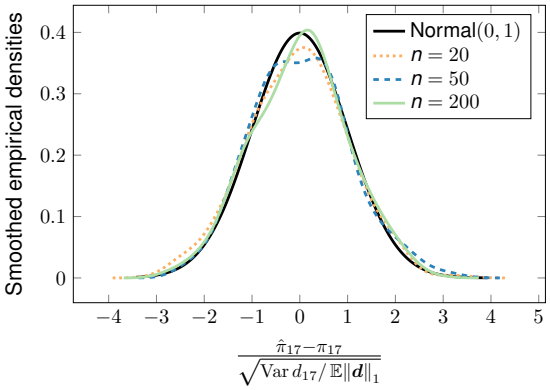


(c) Plug-in: Smoothed empirical densities

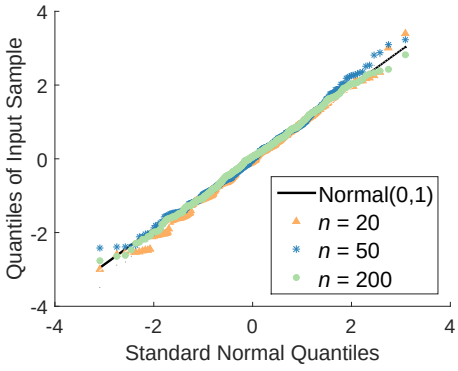


(d) Plug-in: Q-Q plot

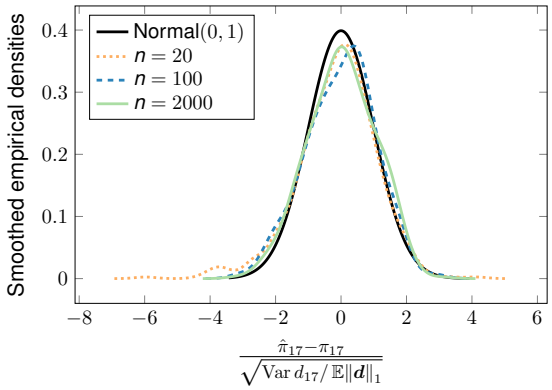
Figure B.1: Illustration of Theorem 3.2.1: The large-sample behavior of the estimator of the centrality of node 5; simulated from power law networks with $\mathbb{E}\, A_{ij} = 0.81 \cdot (ij)^{-0.2}$.



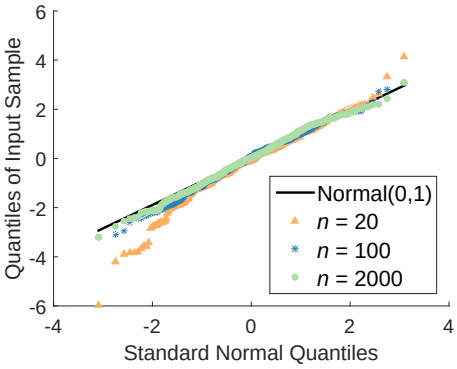
(a) Smoothed empirical densities



(b) Q-Q plot



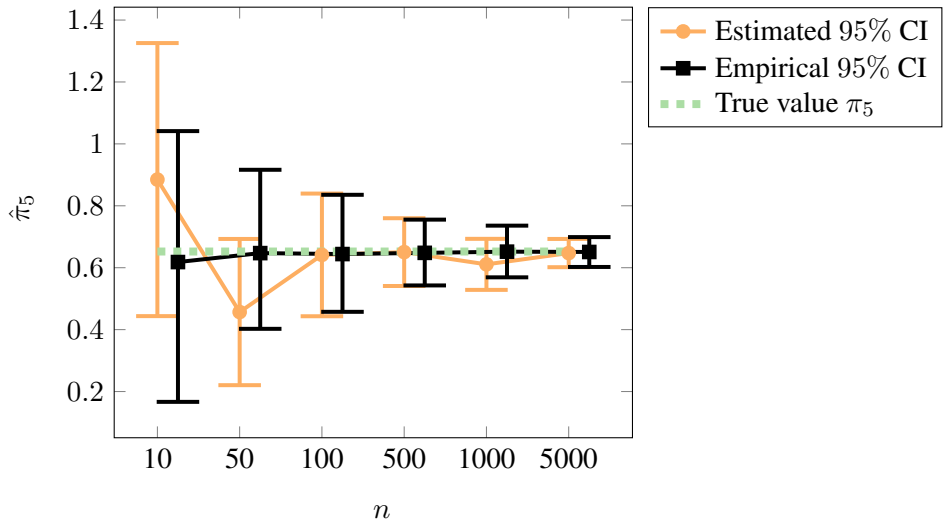
(c) Plug-in: Smoothed empirical densities



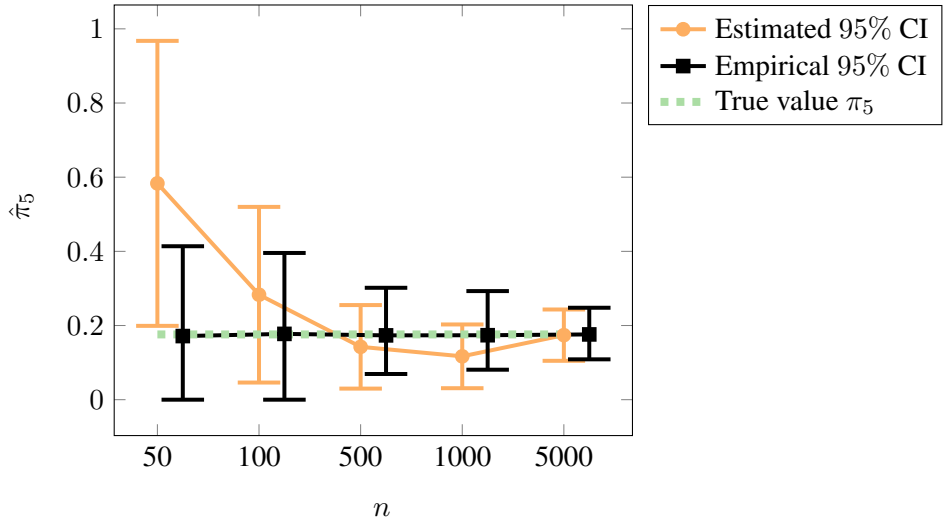
(d) Plug-in: Q-Q plot

Figure B.2: Illustration of Theorem 3.2.1: The large-sample behavior of the estimator of the centrality of node 17; simulated from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.2}$.

B.2.2 Simulations illustrating Theorem 3.2.2



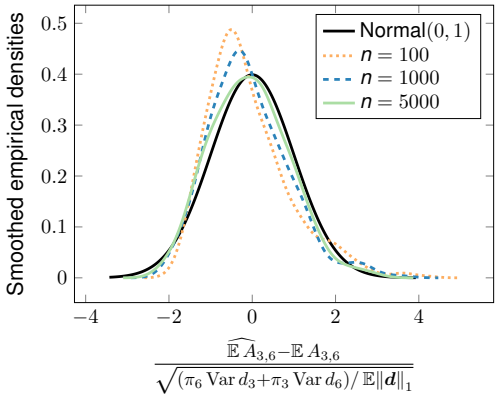
(a) Node 5, $\pi_i = 0.9\ i^{-0.2}$



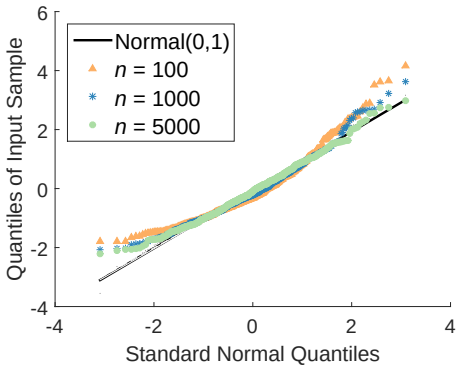
(b) Node 31, $\pi_i = 0.98\ i^{-0.5}$

Figure B.3: The estimator $\hat{\pi}_i$, shown along with its estimated and empirical large-sample confidence intervals; simulated from power law networks.

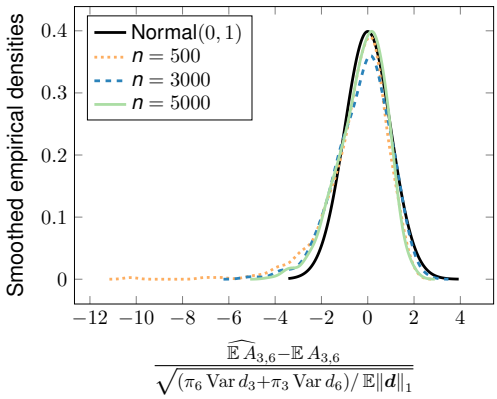
B.2.3 Simulations illustrating Corollary 3.3.2



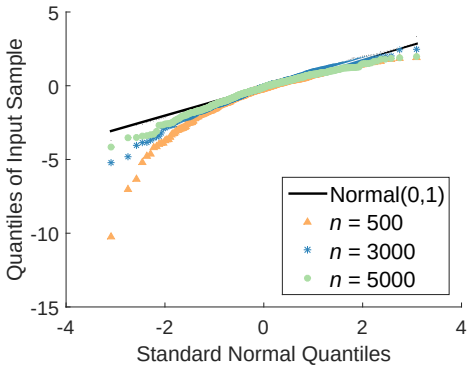
(a) Smoothed empirical densities



(b) Q-Q plot

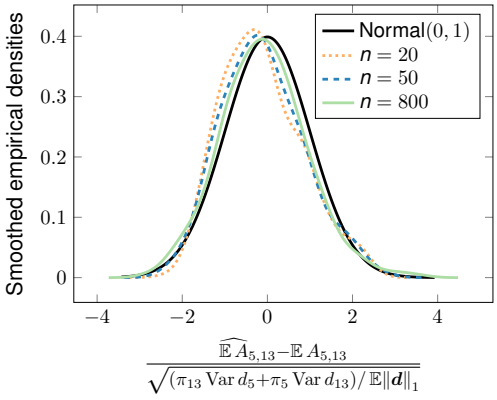


(c) Plug-in: Smoothed empirical densities

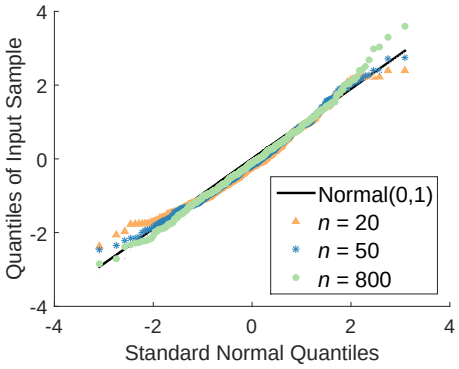


(d) Plug-in: Q-Q plot

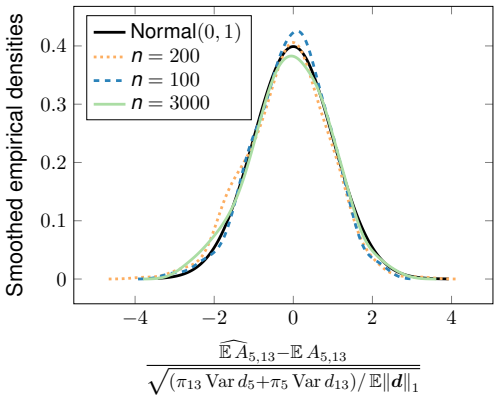
Figure B.4: Illustration of Corollary 3.3.2: The large-sample behavior of the estimator of an edge expectation $\mathbb{E}\, A_{ij}$ between nodes 3 and 6; simulated from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.6}$.



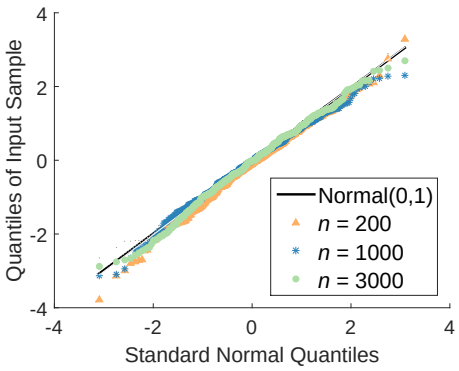
(a) Smoothed empirical densities



(b) Q-Q plot



(c) Plug-in: Smoothed empirical densities



(d) Plug-in: Q-Q plot

Figure B.5: Illustration of Corollary 3.3.2: The large-sample behavior of the estimator of an edge expectation $\mathbb{E}\,A_{ij}$ between nodes 5 and 13; simulated from power law networks with $\mathbb{E}\,A_{ij} = (ij)^{-0.2}$.

Appendix C

Supporting material for Chapter 4

C.1 Lemmas for the proofs in Chapter 4

C.1.1 Approximation of the bias of modularity

We state in the main text that the shift of modularity b in Theorem 4.2.2 Eq. (4.4) is equal to the approximate bias b' to leading order; with

$$b' = \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \right) \delta_{g(i)=g(j)}.$$

More formally, we obtain the following Lemma.

Lemma C.1.1. *Consider Assumptions 1 and 2 and b as defined in Eq.* (4.4)*. Then the following identity holds:*

$$b = \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \left[1 + \mathcal{O}\left(\frac{1}{n^{3/2}} \right) \right] \right) \delta_{g(i)=g(j)}.$$

Proof. Recall from Theorem 4.2.2 Eq. (4.4) that

$$\begin{aligned}
b &= \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} \left(\mathbb{E}\, d_i + \mathbb{E}\, d_j - \|\boldsymbol{\pi}\|_2^2 \right)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \frac{\pi_i^2 \pi_j (\|\boldsymbol{\pi}\|_1 - \pi_i) + \pi_i \pi_j^2 (\|\boldsymbol{\pi}\|_1 - \pi_j) - \pi_i \pi_j \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \left(\frac{\pi_i \pi_j \|\boldsymbol{\pi}\|_1^2 - \pi_i \pi_j \|\boldsymbol{\pi}\|_1^2 + \operatorname{Var} A_{ij} - \operatorname{Var} A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1} \right. \\
&\qquad \left. + \frac{\pi_i^2 \pi_j (\|\boldsymbol{\pi}\|_1 - \pi_i) + \pi_i \pi_j^2 (\|\boldsymbol{\pi}\|_1 - \pi_j) - \pi_i \pi_j \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right) \delta_{g(i)=g(j)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} \sum_{i<j} \Bigg(\frac{\pi_i \pi_j \Big(\|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2 \Big) + \operatorname{Var} A_{ij} - \operatorname{Var} A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1} \\
&\qquad - \frac{\pi_i \pi_j \|\boldsymbol{\pi}\|_1^2 - \pi_i \pi_j \pi_i \|\boldsymbol{\pi}\|_1 + \pi_i^3 \pi_j - \pi_i \pi_j \pi_j \|\boldsymbol{\pi}\|_1 + \pi_i \pi_j^3}{\mathbb{E}\|\boldsymbol{d}\|_1} \Bigg) \delta_{g(i)=g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \Bigg(\frac{\pi_i \pi_j \Big(\|\boldsymbol{\pi}\|_1^2 - \|\boldsymbol{\pi}\|_2^2 \Big) + \operatorname{Var} A_{ij} - \operatorname{Var} A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1} \\
&\qquad - \frac{\pi_i \pi_j (\|\boldsymbol{\pi}\|_1 - \pi_i)(\|\boldsymbol{\pi}\|_1 - \pi_j) + \pi_i^3 \pi_j + \pi_i \pi_j^3}{\mathbb{E}\|\boldsymbol{d}\|_1} \Bigg) \delta_{g(i)=g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i\, \mathbb{E}\, d_j + \operatorname{Var} A_{ij} + \pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1} \right) \delta_{g(i)=g(j)}.
\end{aligned}
$$

Recall from Eq. (3.4) that $\operatorname{cov}(d_i, d_j) = \operatorname{Var} A_{ij}$ for $i \neq j$. Furthermore, it holds that $\mathbb{E}\, d_i d_j = \mathbb{E}\, d_i\, \mathbb{E}\, d_j + \operatorname{cov}(d_i, d_j)$. Hence,

$$
\begin{aligned}
&= \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j + \pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1} \right) \delta_{g(i)=g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \left[1 + \frac{\pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\mathbb{E}\, d_i d_j} \right] \right) \delta_{g(i)=g(j)}.
\end{aligned}
$$

We now define and analyze the error term:

$$
\begin{aligned}
\epsilon &= \frac{\pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\mathbb{E}\, d_i d_j} \\
&= \frac{\pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\mathbb{E}\, d_i\, \mathbb{E}\, d_j + \operatorname{Var} A_{ij}} \\
&= \frac{\pi_i^3 \pi_j + \pi_i \pi_j^3 - \operatorname{Var} A_{ij}}{\pi_i \pi_j (\|\boldsymbol{\pi}\|_1 - \pi_i)(\|\boldsymbol{\pi}\|_1 - \pi_j) + \operatorname{Var} A_{ij}} \\
&= \Theta\left(\frac{\pi_i^3 \pi_j + \pi_i \pi_j^3 - \pi_i \pi_j}{\pi_i \pi_j \|\boldsymbol{\pi}\|_1^2} \right) \quad \text{(Assumption 1)} \\
&= \Theta\left(\frac{\pi_i^2 + \pi_j^2 - 1}{\|\boldsymbol{\pi}\|_1^2} \right) \\
&= \mathcal{O}\left(\frac{1}{\min\left\{ n^2, \|\boldsymbol{\pi}\|_1^2 \right\}} \right) \quad \text{(Assumption 1)} \\
&= \mathcal{O}\left(\frac{1}{n^{3/2}} \right). \quad \text{(Assumption 2)}
\end{aligned}
$$

The required result follows; i.e.,

$$
b = \sum_{j=1}^{n} \sum_{i<j} \left(\mathbb{E}\, A_{ij} - \frac{\mathbb{E}\, d_i d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \left[1 + \mathcal{O}\left(\frac{1}{n^{3/2}} \right) \right] \right) \delta_{g(i)=g(j)}. \quad \text{(Assumption 2)}
$$

□

C.1.2 Lemmas for the proof of Theorem 4.2.1

Lemma C.1.2. *Consider Assumptions 1–4. Then, it holds that*

$$\widehat{Q} = \frac{1}{2}\sum_{j=1}^{n} d_j^w + \sum_{j=1}^{n}\sum_{i<j} \pi_i\pi_j\delta_{g(i)=g(j)} - \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon),$$

with ϵ as defined in Eq. (4.7)*:*

$$\epsilon = \frac{\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)}}{\min(n, \|\boldsymbol{\pi}\|_1)\ \min_l \sqrt{\mathbb{E}\, d_l}}. \tag{C.1}$$

Proof. Recall from Eq. (4.8) that we may write modularity as

$$\widehat{Q} = \sum_{j=1}^{n}\sum_{i<j} A_{ij}\delta_{g(i)=g(j)} - \sum_{j=1}^{n}\sum_{i<j} \widehat{\mathbb{E}\, A}_{ij}\delta_{g(i)=g(j)}. \tag{C.2}$$

Recall from Eq. (3.20) that

$$\begin{aligned}\widehat{\mathbb{E}\, A}_{ij} &= \hat{\pi}_i\hat{\pi}_j \\ &= \pi_i\pi_j + \pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j) + (\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j),\end{aligned}$$

and from Lemma B.1.7 in Appendix B.1.4 that, given Assumptions 1, 2, and 4, it holds that

$$\frac{(\hat{\pi}_i - \pi_i)(\hat{\pi}_j - \pi_j)}{\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)} = \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_i} + \sqrt{\mathbb{E}\, d_j}}\right).$$

As a consequence, we may combine these two results to write

$$\widehat{\mathbb{E}\, A}_{ij} = \pi_i\pi_j + [\pi_j(\hat{\pi}_i - \pi_i) + \pi_i(\hat{\pi}_j - \pi_j)] \cdot \left(1 + \mathcal{O}_P\left(\frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}}\right)\right). \tag{C.3}$$

Focusing on the rightmost sum in Eq. (C.2), we then obtain from Eq. (C.3)

$$\begin{aligned}&\sum_{j=1}^{n}\sum_{i<j} \widehat{\mathbb{E}\, A}_{ij}\delta_{g(i)=g(j)} - \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} \\ &\quad= \left[\sum_{j=1}^{n}\sum_{i<j}\pi_j(\hat{\pi}_i - \pi_i)\delta_{g(i)=g(j)} + \sum_{j=1}^{n}\sum_{i<j}\pi_i(\hat{\pi}_j - \pi_j)\delta_{g(i)=g(j)}\right] \\ &\qquad\cdot \left(1 + \mathcal{O}_P\left(\frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}}\right)\right).\end{aligned}$$

Renaming the indices in the first summand from i to j and vice versa leads to

$$= \left[\sum_{j=1}^{n}\sum_{i\neq j}\pi_i(\hat{\pi}_j - \pi_j)\delta_{g(i)=g(j)}\right] \cdot \left(1 + \mathcal{O}_P\left(\frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}}\right)\right).$$

Hence, $\sum_{j=1}^{n}\sum_{i<j}\widehat{\mathbb{E}\,A_{ij}}\delta_{g(i)=g(j)}$ can be substituted into Eq. (C.2) as follows:

$$\widehat{Q} = \sum_{j=1}^{n}\sum_{i<j} A_{ij}\delta_{g(i)=g(j)} - \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} - \sum_{j=1}^{n}\sum_{i\neq j}\pi_i(\hat{\pi}_j-\pi_j)\delta_{g(i)=g(j)}\cdot\left(1+\mathcal{O}_P\left(\frac{1}{\min_l\sqrt{\mathbb{E}\,d_l}}\right)\right).$$

We now change from a relative error term to an absolute error. In addition, we substitute $\sum_{j=1}^{n}\sum_{i<j}A_{ij}\delta_{g(i)=g(j)} = \frac{1}{2}\sum_{j=1}^{n}d_j^w$, $\hat{\pi}_j = d_j/\sqrt{\|\boldsymbol{d}\|_1}$ and $\sum_{i\neq j}\pi_i\delta_{g(i)=g(j)} = \|\boldsymbol{\pi}\|_1^{g(j),j}$:

$$= \frac{1}{2}\sum_{j=1}^{n}d_j^w - \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} - \left[\sum_{j=1}^{n}\|\boldsymbol{\pi}\|_1^{g(j),j}\left(\frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}}\right) - \sum_{j=1}^{n}\sum_{i\neq j}\pi_i\pi_j\delta_{g(i)=g(j)}\right] + \mathcal{O}_P\left(\frac{1}{\min_l\sqrt{\mathbb{E}\,d_l}}\sum_{j=1}^{n}\|\boldsymbol{\pi}\|_1^{g(j),j}(\hat{\pi}_j-\pi_j)\right).$$

We will treat all error terms involved in Theorem 4.2.1 in Lemma C.1.4 below. To be more precise, we will show that under Assumptions 1–4 it holds

$$\frac{1}{\min_l\sqrt{\mathbb{E}\,d_l}}\sum_{j=1}^{n}\|\boldsymbol{\pi}\|_1^{g(j),j}(\hat{\pi}_j-\pi_j) = \mathcal{O}_P(\epsilon), \tag{C.4}$$

where ϵ is the error term defined in Eq. (C.1). Thus, we obtain the required result

$$\widehat{Q} = \frac{1}{2}\sum_{j=1}^{n}d_j^w + \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} - \sum_{j=1}^{n}\|\boldsymbol{\pi}\|_1^{g(j),j}\frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon).$$

□

Lemma C.1.3. *Consider Assumptions 1–4, and assume Eq.* (4.20) *holds. Then, all non-random terms in modularity in Eq.* (4.20) *may be summed to*

$$b + \mathcal{O}(\epsilon);$$

where b *is defined as in Eq.* (4.4)*:*

$$b = \sum_{j=1}^{n}\sum_{i<j}\frac{\mathbb{E}\,A_{ij}\left(\mathbb{E}\,d_i + \mathbb{E}\,d_j - \|\boldsymbol{\pi}\|_2^2\right)}{\mathbb{E}\|\boldsymbol{d}\|_1}\delta_{g(i)=g(j)}. \tag{C.5}$$

Proof. Recall from Eq. (4.20) that it holds that

$$\begin{aligned}\widehat{Q} = &\sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right] + \sum_{j=1}^{n} \alpha_j^* \,\mathbb{E}\, d_j^w + \sum_{j=1}^{n} \beta_j^* \,\mathbb{E}\, d_j^b \\ &+ \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} + \mathcal{O}_P(\epsilon).\end{aligned} \tag{C.6}$$

We now address the non-random terms in modularity. We treat the non-random terms in the two lines of Eq. (C.6) separately; i.e.,

a) $\sum_{j=1}^{n} \alpha_j^* \,\mathbb{E}\, d_j^w + \sum_{j=1}^{n} \beta_j^* \,\mathbb{E}\, d_j^b$;

b) $\sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}$.

Term a) :

From the definition of α_j^* and β_j^*, we obtain

$$\begin{aligned} a) &= \frac{1}{2} \sum_{j=1}^{n} \mathbb{E}\, d_j^w + \sum_{j=1}^{n} \beta_j^* \,\mathbb{E}\, d_j \\ &= \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} + \sum_{j=1}^{n} \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j}\right] \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}} \\ &= \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} + \left[\frac{1}{2} \sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}}\right] \frac{\sum_{j=1}^{n} \mathbb{E}\, d_j}{\mathbb{E}\|\boldsymbol{d}\|_1} \\ &\quad - \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}} \\ &= \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\,\|\boldsymbol{d}\|_1}} \\ &= b). \end{aligned} \tag{C.7}$$

Term b) :

Via straightforward calculations, one can show that

$$\begin{aligned} b) &= \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} - \frac{1}{2} \sum_{j=1}^{n} \sum_{i \neq j} \frac{\pi_i \left(\pi_j \|\boldsymbol{\pi}\|_1 - \pi_j^2\right)}{\|\boldsymbol{\pi}\|_1 \sqrt{1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}}} \delta_{g(i)=g(j)} \\ &= \sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} \\ &\quad - \frac{1}{2} \sum_{j=1}^{n} \sum_{i \neq j} \left[\pi_i \pi_j - \frac{\pi_i \pi_j^2}{\|\boldsymbol{\pi}\|_1}\right] \left(1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^{-\frac{1}{2}} \delta_{g(i)=g(j)}. \end{aligned} \tag{C.8}$$

We know from Eq. (B.3) that from Assumption 1 it follows that $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$. As a consequence, we can apply a convergent Taylor expansion to $f(x) = (1-x)^{-1/2}$ at 0 to obtain

$$\left(1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^{-\frac{1}{2}} = 1 + \frac{1}{2}\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} + \mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]. \tag{C.9}$$

As a consequence, it follows that we may express Eq. (C.8) as

$$\begin{aligned} b) = &\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} - \frac{1}{2}\sum_{j=1}^{n}\sum_{i\neq j}\pi_i\pi_j\delta_{g(i)=g(j)} \\ &- \sum_{j=1}^{n}\sum_{i<j}\left[\frac{1}{2}\pi_i\pi_j\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} + \pi_i\pi_j\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)} \qquad \text{(C.10)} \\ &+ \frac{1}{2}\sum_{j=1}^{n}\sum_{i\neq j}\left[\frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1} + \frac{1}{2}\frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} + \frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)}. \qquad \text{(C.11)} \end{aligned}$$

We identify the first terms in Eqs. (C.10) and (C.11) as the terms of leading order. We will show in Lemma C.1.4 below that under Assumptions 1–4 the remaining terms satisfy

$$\begin{aligned} &- \sum_{j=1}^{n}\sum_{i<j}\left[\pi_i\pi_j\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)} \\ &\quad + \frac{1}{2}\sum_{j=1}^{n}\sum_{i\neq j}\left[\frac{1}{2}\frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} + \frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)} \qquad \text{(C.12)} \\ &= \mathcal{O}(\epsilon), \end{aligned}$$

where we remind the reader that ϵ is the error term defined in Eq. (4.7).

Finally, considering the leading-order terms in (C.10) and (C.11), it then follows from the identity

$$\sum_{j=1}^{n}\sum_{i\neq j}\pi_i\pi_j^2\delta_{g(i)=g(j)} = \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j(\pi_i+\pi_j)\delta_{g(i)=g(j)}$$

that

$$b) = \frac{1}{2}\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\left[\frac{\pi_i+\pi_j}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]\delta_{g(i)=g(j)} + \mathcal{O}(\epsilon). \tag{C.13}$$

We may then combine terms a) and b) using Eqs. (C.7) and (C.13), whence

$$a) + b) = \sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\left[\frac{\pi_i+\pi_j}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]\delta_{g(i)=g(j)} + \mathcal{O}(\epsilon).$$

In order to gain interpretability, we rearrange the term $a) + b)$ even further:

$$
\begin{aligned}
&= \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right] \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon) \\
&= \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right] \left[1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right] \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon) \\
&= \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right] \delta_{g(i)=g(j)} \\
&\quad - \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right] \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} \left(\mathbb{E}\, d_i + \mathbb{E}\, d_j - \|\boldsymbol{\pi}\|_2^2 \right)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} \\
&\quad + \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} (\pi_i + \pi_j)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} \\
&\quad - \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right] \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \delta_{g(i)=g(j)} + \mathcal{O}(\epsilon).
\end{aligned}
\tag{C.14}
$$

We will show in Lemma C.1.4 below that under Assumptions 1–4 it holds

$$
\begin{aligned}
&\sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} (\pi_i + \pi_j)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)} \\
&\quad - \sum_{j=1}^{n} \sum_{i<j} \mathbb{E}\, A_{ij} \left[\frac{\pi_i \|\boldsymbol{\pi}\|_1 + \pi_j \|\boldsymbol{\pi}\|_1 - \|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1} \right] \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \delta_{g(i)=g(j)} \\
&= \mathcal{O}(\epsilon).
\end{aligned}
\tag{C.15}
$$

Recalling from Eq. (C.5) that b is defined as

$$
b = \sum_{j=1}^{n} \sum_{i<j} \frac{\mathbb{E}\, A_{ij} \left(\mathbb{E}\, d_i + \mathbb{E}\, d_j - \|\boldsymbol{\pi}\|_2^2 \right)}{\mathbb{E}\|\boldsymbol{d}\|_1} \delta_{g(i)=g(j)}.
$$

we conclude from Eqs. (C.14) and (C.15) the required result:

$$
a) + b) = b + \mathcal{O}(\epsilon).
$$

□

Lemma C.1.4. *Consider Assumptions 1–4. Then, the five error terms from Eqs.* (C.4)*,* (4.13)*,* (4.16)*,* (C.12)*, and* (C.15) *are* $\mathcal{O}(\epsilon)$ *with* ϵ *defined as in Eq.* (4.7)*:*

$$
\epsilon = \frac{\sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)}}{\min(n, \|\boldsymbol{\pi}\|_1) \, \min_l \sqrt{\mathbb{E}\, d_l}}.
$$

Proof. We now define and address the five error terms cited in proof of Theorem 4.2.1; we call these $\epsilon^{(1)}, \epsilon^{(2)}, \ldots, \epsilon^{(5)}$.

Term $\epsilon^{(1)}$: Recalling Eq. (C.4), we define

$$\begin{aligned}\epsilon^{(1)} &= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} (\hat{\pi}_j - \pi_j) \\ &= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(\frac{d_j}{\sqrt{\|\boldsymbol{d}\|_1}} - \pi_j \right).\end{aligned}$$

First, we apply a Taylor expansion to $(\|\boldsymbol{d}\|_1 / \mathbb{E}\|\boldsymbol{d}\|_1)^{-1/2} = f(x) = x^{-1/2}$ at 1, leading to

$$\frac{1}{\sqrt{\|\boldsymbol{d}\|_1}} = \frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \left[1 + \mathcal{O}_P\left(\sqrt{\frac{\mathrm{Var}\|\boldsymbol{d}\|_1}{(\mathbb{E}\|\boldsymbol{d}\|_1)^2}} \right) \right],$$

and then control the remainder using Chebyshev's inequality. As a consequence, we obtain from Assumption 4 ($\mathrm{Var}\, A_{ij} = \Theta(\mathbb{E}\, A_{ij})$) that

$$\epsilon^{(1)} = \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(\frac{d_j \left[1 + \mathcal{O}_P\big(1/\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}\big) \right]}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} - \pi_j \right).$$

From Chebyshev's inequality and Assumptions 2 and 4, we know that $d_j = \mathbb{E}\, d_j + \mathcal{O}_P\big(\sqrt{\mathbb{E}\, d_j}\big) = \mathbb{E}\, d_j \big[1 + \mathcal{O}_P(1/\sqrt{\mathbb{E}\, d_j}) \big]$. It follows that

$$\begin{aligned}&= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(\frac{\mathbb{E}\, d_j \left[1 + \mathcal{O}_P\big(1/\sqrt{\mathbb{E}\, d_j}\big) \right]}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} - \pi_j \right) \\ &= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(\pi_j \frac{\left[1 + \mathcal{O}_P\big(1/\sqrt{\mathbb{E}\, d_j}\big) \right] \left[1 - \pi_j / \|\boldsymbol{\pi}\|_1 \right]}{\left[1 - \|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2 \right]^{1/2}} - \pi_j \right).\end{aligned}$$

Since $\|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$ (Eq. (B.3), following from Assumption 1), we can apply a convergent Taylor expansion to $f(x) = (1-x)^{-1/2}$ at 0 (as in Eq. (C.9)). Furthermore, the remainder term $\left(\|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2 \right)^2$ in this Taylor expansion satisfies $\left(\|\boldsymbol{\pi}\|_2^2 / \|\boldsymbol{\pi}\|_1^2 \right)^2 = \mathcal{O}\big(1/n^2\big) = \mathcal{O}\big(1/\sqrt{\mathbb{E}\, d_j}\big)$ (Assumptions 1 and 3). Hence, we obtain

$$\begin{aligned}&= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(\pi_j \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_j}} \right) \right] \left[1 - \frac{\pi_j}{\|\boldsymbol{\pi}\|_1} \right] \right. \\ &\qquad\qquad \left. \left[1 + \frac{1}{2} \left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right) \right] - \pi_j \right) \\ &= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \left(-\frac{\pi_j^2}{\|\boldsymbol{\pi}\|_1} + \frac{1}{2} \pi_j \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right) \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_j}} \right) \right] \\ &= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \sum_{i \neq j} \pi_i \pi_j \left(-\frac{\pi_j}{\|\boldsymbol{\pi}\|_1} + \frac{1}{2} \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right) \delta_{g(i)=g(j)} \left[1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_j}} \right) \right]\end{aligned}$$

$$= \frac{1}{\min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \sum_{i \neq j} \pi_i \pi_j \delta_{g(i)=g(j)} \cdot \mathcal{O}_P\left(\frac{1}{n}\right). \quad \text{(Assumption 1, Eq. (B.3))} \tag{C.16}$$

Term $\epsilon^{(2)}$: We now analyze the second error term. Recalling Eq. (4.13), define

$$\epsilon^{(2)} = \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \frac{1}{\mathbb{E}\|\boldsymbol{d}\|_1}.$$

From Chebyshev's inequality and Assumption 4 it follows that

$$= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \frac{1}{\mathbb{E}\|\boldsymbol{d}\|_1} \left(1 + \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\, d_j}}\right)\right). \tag{C.17}$$

This expression is smaller than $\epsilon^{(3)}$ as defined in Eq. (C.18).

Term $\epsilon^{(3)}$: We now analyze the third error term. Recalling Eq. (4.16), define

$$\epsilon^{(3)} = \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \left(\frac{\|\boldsymbol{d}\|_1}{\mathbb{E}\|\boldsymbol{d}\|_1} - 1\right) \frac{1}{\sqrt{\mathbb{E}\, d_j}}.$$

Applying Chebyshev's inequality leads to

$$\begin{aligned}
&= \frac{1}{2} \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \cdot \mathcal{O}_P\left(\frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}}\right) \cdot \frac{1}{\sqrt{\mathbb{E}\, d_j}} \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \mathcal{O}_P\left(\frac{\mathbb{E}\, d_j}{\mathbb{E}\|\boldsymbol{d}\|_1 \sqrt{\mathbb{E}\, d_j}}\right) \qquad \text{(C.18)} \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \mathcal{O}_P\left(\frac{\sqrt{\mathbb{E}\, d_j}}{\mathbb{E}\|\boldsymbol{d}\|_1}\right) \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \pi_j \sqrt{\frac{\pi_j \|\boldsymbol{\pi}\|_1}{\pi_j^2 \|\boldsymbol{\pi}\|_1^4}} \mathcal{O}_P\left(\sqrt{\frac{1 - \pi_j/\|\boldsymbol{\pi}\|_1}{1 - \|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2}}\right) \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \pi_j \sqrt{\frac{\pi_j \|\boldsymbol{\pi}\|_1}{\pi_j^2 \|\boldsymbol{\pi}\|_1^4}} \mathcal{O}_P\left(\sqrt{1 + \frac{1}{n}}\right) \quad \text{(Assumption 1, Eqs. (B.3), (C.9))} \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \pi_j \sqrt{\frac{1}{\pi_j \|\boldsymbol{\pi}\|_1 \|\boldsymbol{\pi}\|_1^2}} \mathcal{O}_P\left(\sqrt{1 + \frac{1}{n}}\right) \\
&= \sum_{j=1}^{n} \|\boldsymbol{\pi}\|_1^{g(j),j} \pi_j \sqrt{\frac{1 - \pi_j/\|\boldsymbol{\pi}\|_1}{\mathbb{E}\, d_j \|\boldsymbol{\pi}\|_1^2}} \mathcal{O}_P\left(\sqrt{1 + \frac{1}{n}}\right) \\
&= \frac{\sum_{j=1} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)}}{\min_l \sqrt{\mathbb{E}\, d_l}\, \|\boldsymbol{\pi}\|_1} \mathcal{O}_P\left(\sqrt{1 + \frac{1}{n}}\right). \quad \text{(Assumption 1)} \qquad \text{(C.19)}
\end{aligned}$$

Term $\epsilon^{(4)}$: We now analyze the fourth error term. Recalling Eq. (C.12), define

$$\begin{aligned}
\epsilon^{(4)} &= -\sum_{j=1}^{n}\sum_{i<j}\left[\pi_i\pi_j\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)} \\
&\quad + \frac{1}{2}\sum_{j=1}^{n}\sum_{i\neq j}\left[\frac{1}{2}\frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} + \frac{\pi_i\pi_j^2}{\|\boldsymbol{\pi}\|_1}\,\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\right]\delta_{g(i)=g(j)} \\
&= -\mathcal{O}\left[\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^2\right]\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)} \\
&\quad + \mathcal{O}\left[\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]\sum_{j=1}^{n}\sum_{i\neq j}\pi_i\pi_j\frac{\pi_j}{\|\boldsymbol{\pi}\|_1}\delta_{g(i)=g(j)} \\
&= \mathcal{O}\left(\frac{1}{n^2}\right)\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)}. \quad \text{(Assumption 1, Eq. (B.3))}
\end{aligned} \tag{C.20}$$

Term $\epsilon^{(5)}$: We now analyze the fifth error term. Recalling Eq. (C.15), define

$$\begin{aligned}
\epsilon^{(5)} =& \sum_{j=1}^{n}\sum_{i<j}\frac{\mathbb{E}\,A_{ij}(\pi_i+\pi_j)}{\mathbb{E}\|\boldsymbol{d}\|_1}\delta_{g(i)=g(j)} \\
&- \sum_{j=1}^{n}\sum_{i<j}\mathbb{E}\,A_{ij}\left[\frac{\pi_i\|\boldsymbol{\pi}\|_1+\pi_j\|\boldsymbol{\pi}\|_1-\|\boldsymbol{\pi}\|_2^2}{\mathbb{E}\|\boldsymbol{d}\|_1}\right]\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\delta_{g(i)=g(j)} \\
=& \sum_{j=1}^{n}\sum_{i<j}\frac{\mathbb{E}\,A_{ij}(\pi_i+\pi_j)}{\mathbb{E}\|\boldsymbol{d}\|_1}\delta_{g(i)=g(j)}\left[1-\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1}\right] \\
&+ \sum_{j=1}^{n}\sum_{i<j}\frac{\mathbb{E}\,A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1}\delta_{g(i)=g(j)}\left(\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1}\right)^2 \\
\leq& \frac{2\max_l\pi_l}{\mathbb{E}\|\boldsymbol{d}\|_1}\sum_{j=1}^{n}\sum_{i<j}\mathbb{E}\,A_{ij}\delta_{g(i)=g(j)}\left[1+\mathcal{O}\left(\max_l\pi_l\right)\right] \\
&+ \sum_{j=1}^{n}\sum_{i<j}\frac{\mathbb{E}\,A_{ij}}{\mathbb{E}\|\boldsymbol{d}\|_1}\delta_{g(i)=g(j)}\left(\max_l\pi_l\right)^2 \quad \text{(Assumption 1, Eq. (B.3))} \\
=& \frac{2\max_l\pi_l+\mathcal{O}\big(\max_l\pi_l^2\big)}{\|\boldsymbol{\pi}\|_1^2}\left[1-\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right]^{-1}\sum_{j=1}^{n}\sum_{i<j}\mathbb{E}\,A_{ij}\delta_{g(i)=g(j)}.
\end{aligned}$$

Applying a convergent Taylor expansion to $f(x) = (1-x)^{-1}$ at 0 with $x = \|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2$ (Assumption 1 and Eq. (B.3)), we obtain

$$\begin{aligned}
=& \frac{2\max_l\pi_l+\mathcal{O}\big(\max_l\pi_l^2\big)}{\|\boldsymbol{\pi}\|_1^2}\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]\sum_{j=1}^{n}\sum_{i<j}\mathbb{E}\,A_{ij}\delta_{g(i)=g(j)} \\
=& \mathcal{O}\left(\frac{1}{n\,\|\boldsymbol{\pi}\|_1}+\frac{1}{n^2}\right)\sum_{j=1}^{n}\sum_{i<j}\pi_i\pi_j\delta_{g(i)=g(j)}. \quad \text{(Assumption 1)}
\end{aligned} \tag{C.21}$$

As a consequence of Eqs. (C.16)–(C.21), we now know that the error terms $\epsilon^{(1)}, \epsilon^{(2)}, \ldots, \epsilon^{(5)}$ in our analysis of modularity satisfy

$$\begin{aligned}
&\epsilon^{(1)} + \epsilon^{(2)} + \epsilon^{(3)} + \epsilon^{(4)} + \epsilon^{(5)} \\
&= \mathcal{O}_P\bigg(\frac{1}{n\,\min_l \sqrt{\mathbb{E}\,d_l}} + \frac{1}{\|\boldsymbol{\pi}\|_1\,\min_l \sqrt{\mathbb{E}\,d_l}} + \frac{1}{n^2} + \frac{1}{\|\boldsymbol{\pi}\|_1 n}\bigg) \sum_{j=1}^{n}\sum_{i<j} \pi_i\pi_j \delta_{g(i)=g(j)}.
\end{aligned}$$

From Assumption 3, it follows that $\min_l \sqrt{\mathbb{E}\,d_l} = o\Big(\sqrt{n^2}\Big) = o(n)$. Hence,

$$= \mathcal{O}_P\bigg(\frac{1}{n\,\min_l \sqrt{\mathbb{E}\,d_l}} + \frac{1}{\|\boldsymbol{\pi}\|_1\,\min_l \sqrt{\mathbb{E}\,d_l}}\bigg) \sum_{j=1}^{n}\sum_{i<j} \pi_i\pi_j \delta_{g(i)=g(j)}.$$

Recall from Eq. (4.7) that

$$\epsilon = \frac{\sum_{j=1}^{n}\sum_{i<j} \pi_i\pi_j\delta_{g(i)=g(j)}}{\min(n, \|\boldsymbol{\pi}\|_1)\,\min_l \sqrt{\mathbb{E}\,d_l}}.$$

As a consequence, we conclude the required result of this lemma:

$$\epsilon^{(1)} + \epsilon^{(2)} + \epsilon^{(3)} + \epsilon^{(4)} + \epsilon^{(5)} = \mathcal{O}_P(\epsilon).$$

□

Lemma C.1.5. *Consider Assumption 1 ($\pi_i/\|\boldsymbol{\pi}\|_1 = \mathcal{O}(1/n)$), and define ϵ as in Lemma C.1.4. Then, the following identity holds:*

$$\begin{aligned}
&\sum_{j=1}^{n} \alpha_j^* \big[d_j^w - \mathbb{E}\,d_j^w\big] + \sum_{j=1}^{n} \beta_j^* \Big[d_j^b - \mathbb{E}\,d_j^b\Big] \\
&= \sum_{j=1}^{n} \alpha_j \big[d_j^w - \mathbb{E}\,d_j^w\big] + \sum_{j=1}^{n} \beta_j \Big[d_j^b - \mathbb{E}\,d_j^b\Big] + \mathcal{O}_P(\epsilon),
\end{aligned}$$

with $\alpha_j = 0.5 + \beta_j$ and

$$\beta_j = \frac{\sum_{l=1}^{n} \mathbb{E}\,d_l^w}{2\,\mathbb{E}\|\boldsymbol{d}\|_1} - \frac{\mathbb{E}\,d_j^w}{\mathbb{E}\,d_j}.$$

Proof. We first address how β_j and β_j^* relate:

$$\begin{aligned}
\beta_j^* &= \left[\frac{1}{2}\sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\,d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j}\right] \frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \\
&= \left[\frac{\sum_{l=1}^{n}\sum_{m<l} \mathbb{E}\,A_{lm}\delta_{g(l)=g(m)}}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \frac{\|\boldsymbol{\pi}\|_1(1-\pi_l/\|\boldsymbol{\pi}\|_1)}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} - \|\boldsymbol{\pi}\|_1^{g(j),j}\right] \frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \\
&= \left[\frac{\sum_{l=1}^{n}\sum_{m<l} \mathbb{E}\,A_{lm}\delta_{g(l)=g(m)}}{2\sum_{l=1}^{n}\sum_{m<l} \mathbb{E}\,A_{lm}} \frac{\|\boldsymbol{\pi}\|_1(1-\pi_l/\|\boldsymbol{\pi}\|_1)}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_1^{g(j),j}}{\|\boldsymbol{\pi}\|_1}\right] \\
&\quad \cdot \frac{1}{\sqrt{1 - \|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2}}
\end{aligned}$$

From Assumption 1 and Eq. (B.2), we know that $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 \leq \max_i \pi_i \|\boldsymbol{\pi}\|_1/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$. Hence, we can apply a convergent Taylor expansion to $f(x) = (1-x)^{-1/2}$ at $x = 0$. We obtain

$$
\begin{aligned}
&= \left[\frac{\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}\delta_{g(l)=g(m)}}{2\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}} - \frac{\|\boldsymbol{\pi}\|_1^{g(j),j}}{\|\boldsymbol{\pi}\|_1}\right]\left[1+\mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1}\right)\right] \\
&= \left[\frac{\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}\delta_{g(l)=g(m)}}{2\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}} - \frac{\pi_j\|\boldsymbol{\pi}\|_1^{g(j),j}}{\pi_j\|\boldsymbol{\pi}\|_1(1-\pi_j/\|\boldsymbol{\pi}\|_1)}\right]\left(1-\frac{\pi_j}{\|\boldsymbol{\pi}\|_1}\right) \\
&\quad \cdot \left[1+\mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1}\right)\right] \\
&= \left[\frac{\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}\delta_{g(l)=g(m)}}{2\sum_{l=1}^n \sum_{m<l} \mathbb{E}\, A_{lm}} - \frac{\pi_j\|\boldsymbol{\pi}\|_1^{g(j),j}}{\pi_j\|\boldsymbol{\pi}\|_1(1-\pi_j/\|\boldsymbol{\pi}\|_1)}\right]\left[1+\mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1}\right)\right] \\
&= \left[\frac{\sum_{l=1}^n \mathbb{E}\, d_l^w}{2\,\mathbb{E}\|\boldsymbol{d}\|_1} - \frac{\mathbb{E}\, d_j^w}{\mathbb{E}\, d_j}\right]\left[1+\mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1}\right)\right] \\
&= \left[\frac{\sum_{l=1}^n \mathbb{E}\, d_l^w}{2\,\mathbb{E}\|\boldsymbol{d}\|_1} - \frac{\mathbb{E}\, d_j^w}{\mathbb{E}\, d_j}\right]\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right] \quad \text{(Assumption 1).} \\
&= \beta_j\left[1+\mathcal{O}\left(\frac{1}{n}\right)\right]. \qquad \text{(C.22)}
\end{aligned}
$$

We now address the error introduced into the decomposition of modularity $\widehat{Q} - b$ in Eq. (4.23) by changing from β_j^* to β_j. Substituting the result in Eq. (C.22) into Eq. (4.23), we obtain

$$
\begin{aligned}
&\sum_{j=1}^n \alpha_j^*\left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^n \beta_j^*\left[d_j^b - \mathbb{E}\, d_j^b\right] \\
&= \sum_{j=1}^n (0.5+\beta_j)\left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^n \beta_j\left[d_j^b - \mathbb{E}\, d_j^b\right] + \mathcal{O}\left(\frac{1}{n}\sum_{j=1}^n \beta_j[d_j - \mathbb{E}\, d_j]\right).
\end{aligned}
$$

We now address the error term:

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^n \beta_j[d_j - \mathbb{E}\, d_j] &= \frac{1}{n}\sum_{j=1}^n \beta_j \mathcal{O}_P\left(\sqrt{\mathbb{E}\, d_j}\right) \quad \text{(Chenyshev's inequality)} \\
&= \mathcal{O}_P\left(\frac{1}{n}\sum_{j=1}^n\left(\frac{\sum_{l=1}^n \mathbb{E}\, d_l^w}{2\,\mathbb{E}\|\boldsymbol{d}\|_1} + \frac{\mathbb{E}\, d_j^w}{\mathbb{E}\, d_j}\right)\sqrt{\mathbb{E}\, d_j}\right) \\
&= \mathcal{O}_P\left(\frac{1}{n}\sum_{j=1}^n\left(\frac{\sum_{l=1}^n \mathbb{E}\, d_l^w}{2\,\mathbb{E}\|\boldsymbol{d}\|_1}\frac{\mathbb{E}\, d_j}{\sqrt{\mathbb{E}\, d_j}} + \frac{\mathbb{E}\, d_j^w}{\sqrt{\mathbb{E}\, d_j}}\right)\right) \\
&= \mathcal{O}_P\left(\frac{1}{n\min_l \sqrt{\mathbb{E}\, d_l}}\left(\frac{\sum_{l=1}^n \mathbb{E}\, d_l^w \sum_{j=1}^n \mathbb{E}\, d_j}{2\,\mathbb{E}\|\boldsymbol{d}\|_1} + \sum_{j=1}^n \mathbb{E}\, d_j^w\right)\right)
\end{aligned}
$$

$$= \mathcal{O}_P\left(\frac{1}{n \min_l \sqrt{\mathbb{E}\, d_l}} \sum_{j=1}^{n} \sum_{i \neq j} \pi_i \pi_j \delta_{g(i)=g(j)}\right)$$
$$= \mathcal{O}_P(\epsilon).$$

As a consequence, we conclude the required result of this lemma; i.e.,

$$\sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right]$$
$$= \sum_{j=1}^{n} \alpha_j \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j \left[d_j^b - \mathbb{E}\, d_j^b\right] + \mathcal{O}_P(\epsilon).$$

□

C.1.3 Lemmas for the proof of Theorem 4.2.2

Lemma C.1.6. *It holds that*

$$X_n = \sum_{j=1}^{n} \sum_{i<j} c_{ij}[A_{ij} - \mathbb{E}\, A_{ij}],$$
$$c_{ij} = \delta_{g(i)=g(j)} + \beta_i^* + \beta_j^*.$$

Proof. We write X_n as a sum of independent, zero-mean random variables:

$$\begin{aligned}
X_n &= \sum_{j=1}^{n} \alpha_j^* \left[d_j^w - \mathbb{E}\, d_j^w\right] + \sum_{j=1}^{n} \beta_j^* \left[d_j^b - \mathbb{E}\, d_j^b\right] \\
&= \sum_{j=1}^{n} \sum_{i \neq j} \alpha_j^* [A_{ij} - \mathbb{E}\, A_{ij}] \delta_{g(i)=g(j)} + \sum_{j=1}^{n} \sum_{i \neq j} \beta_j^* [A_{ij} - \mathbb{E}\, A_{ij}] \delta_{g(i)\neq g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \left(\alpha_i^* + \alpha_j^*\right) [A_{ij} - \mathbb{E}\, A_{ij}] \delta_{g(i)=g(j)} \\
&\quad + \sum_{j=1}^{n} \sum_{i<j} \left(\beta_i^* + \beta_j^*\right) [A_{ij} - \mathbb{E}\, A_{ij}] \delta_{g(i)\neq g(j)} \\
&= \sum_{j=1}^{n} \sum_{i<j} \left[\left(\alpha_i^* + \alpha_j^*\right) \delta_{g(i)=g(j)} + \left(\beta_i^* + \beta_j^*\right) \delta_{g(i)\neq g(j)}\right] [A_{ij} - \mathbb{E}\, A_{ij}] \\
&= \sum_{j=1}^{n} \sum_{i<j} \left[\left(1 + \beta_i^* + \beta_j^*\right) \delta_{g(i)=g(j)} + \left(\beta_i^* + \beta_j^*\right) \delta_{g(i)\neq g(j)}\right] [A_{ij} - \mathbb{E}\, A_{ij}] \\
&= \sum_{j=1}^{n} \sum_{i<j} \underbrace{\left[\delta_{g(i)=g(j)} + \beta_i^* + \beta_j^*\right]}_{c_{ij}} [A_{ij} - \mathbb{E}\, A_{ij}] \\
&= \sum_{j=1}^{n} \sum_{i<j} c_{ij} [A_{ij} - \mathbb{E}\, A_{ij}].
\end{aligned}$$

□

Lemma C.1.7. *Consider Assumption 1. Then it holds both that*

$$c_{ij} = \mathcal{O}(1),$$

and that c_{ij} may be expressed as a function only of the group assignments $g(i) = m$ and $g(j) = l$:

$$c_{lm} = \delta_{l=m} + \sum_{k=1}^{K} \left(\frac{\|\boldsymbol{\pi}\|_1^{k,\emptyset}}{\|\boldsymbol{\pi}\|_1} \right)^2 - \frac{\|\boldsymbol{\pi}\|_1^{l,\emptyset}}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_1^{m,\emptyset}}{\|\boldsymbol{\pi}\|_1} + \mathcal{O}\left(\frac{1}{n}\right).$$

Proof. From Eq. (4.26) and the definitions of α^*, β^* in Eq. (4.19), we see that

$$\begin{aligned} & c_{ij} - \delta_{g(i)=g(j)} \\ & \quad = \left[\sum_{l=1}^{n} \|\boldsymbol{\pi}\|_1^{g(l),l} \frac{\mathbb{E}\, d_l}{\mathbb{E}\|\boldsymbol{d}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} - \|\boldsymbol{\pi}\|_1^{g(i),i} \right] \frac{1}{\sqrt{\mathbb{E}\|\boldsymbol{d}\|_1}} \end{aligned}$$

$$= \left[\sum_{l=1}^{n} \frac{\left(\|\boldsymbol{\pi}\|_1^{g(l),\emptyset} - \pi_l \right) \pi_l}{\|\boldsymbol{\pi}\|_1} \left(\frac{1 - \frac{\pi_l}{\|\boldsymbol{\pi}\|_1}}{1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}} \right) - \|\boldsymbol{\pi}\|_1^{g(j),j} - \|\boldsymbol{\pi}\|_1^{g(i),i} \right] \frac{\left(1 - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right)^{-\frac{1}{2}}}{\|\boldsymbol{\pi}\|_1}.$$

From Assumption 1 and Eq. (B.2), we know that $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 \leq \max_i \pi_i \|\boldsymbol{\pi}\|_1/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$. Hence, we can apply a convergent Taylor expansion to $f(x) = (1-x)^{-\alpha}, \alpha = 1/2, 1$ at $x = 0$. We obtain

$$= \left[\frac{\sum_{k=1}^{K} \left(\|\boldsymbol{\pi}\|_1^{k,\emptyset} \right)^2 - \|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),j} - \|\boldsymbol{\pi}\|_1^{g(i),i} \right] \frac{\left[1 + \mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1} \right) \right]}{\|\boldsymbol{\pi}\|_1} \tag{C.23}$$

$$\begin{aligned} &= \left[\frac{\sum_{k=1}^{K} \left(\|\boldsymbol{\pi}\|_1^{k,\emptyset} \right)^2}{\|\boldsymbol{\pi}\|_1} - \|\boldsymbol{\pi}\|_1^{g(j),\emptyset} - \|\boldsymbol{\pi}\|_1^{g(i),\emptyset} \right] \frac{\left[1 + \mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1} \right) \right]}{\|\boldsymbol{\pi}\|_1} \\ &\quad + \left[\frac{\pi_j}{\|\boldsymbol{\pi}\|_1} + \frac{\pi_i}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2} \right] \left[1 + \mathcal{O}\left(\frac{\max_i \pi_i}{\|\boldsymbol{\pi}\|_1} \right) \right]. \end{aligned}$$

Since $\|\boldsymbol{\pi}\|_2^2/\|\boldsymbol{\pi}\|_1^2 \leq \max_i \pi_i \|\boldsymbol{\pi}\|_1/\|\boldsymbol{\pi}\|_1^2 = \mathcal{O}(1/n)$, it follows further that

$$= \left[\frac{\sum_{k=1}^{K} \left(\|\boldsymbol{\pi}\|_1^{k,\emptyset} \right)^2}{\|\boldsymbol{\pi}\|_1^2} - \frac{\|\boldsymbol{\pi}\|_1^{g(j),\emptyset}}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_1^{g(i),\emptyset}}{\|\boldsymbol{\pi}\|_1} \right] \left[1 + \mathcal{O}\left(\frac{1}{n} \right) \right] + \mathcal{O}\left(\frac{1}{n} \right). \tag{C.24}$$

The first term in Eq. (C.24) is $\mathcal{O}(1)$, and thus we conclude the first of the two required results of this lemma: $c_{ij} = \mathcal{O}(1)$. This in turn allows us to combine the relative and additive error terms.

Furthermore we see that c_{ij} is, up to an additive error term of order at most $1/n$, a function only of $g(i)$ and $g(j)$. This leads to the second of the two required results of this lemma:

$$c_{ij} = \delta_{g(i)=g(j)} + \sum_{k=1}^{K} \left(\frac{\|\boldsymbol{\pi}\|_1^{k,\emptyset}}{\|\boldsymbol{\pi}\|_1} \right)^2 - \frac{\|\boldsymbol{\pi}\|_1^{g(i),\emptyset}}{\|\boldsymbol{\pi}\|_1} - \frac{\|\boldsymbol{\pi}\|_1^{g(j),\emptyset}}{\|\boldsymbol{\pi}\|_1} + \mathcal{O}\left(\frac{1}{n}\right).$$

□

Lemma C.1.8. *Denote ϵ the error term defined in Eq.* (4.7) *and X_n the sequence of random variables from Eq.* (4.25)*. Consider Assumptions 1–3. Then, it holds that*

$$(\operatorname{Var} X_n)^{-\frac{1}{2}} \epsilon \xrightarrow{n} 0.$$

Proof. As in Lemma 4.28, we first define

$$a_k = \|\boldsymbol{\pi}\|_1^{k,\emptyset} = \sum_{i=1}^{n} \pi_i \delta_{g(i)=k},$$

whence

$$\sum_{j=1}^{n} \sum_{i<j} \pi_i \pi_j \delta_{g(i)=g(j)} \leq \frac{1}{2} \|\boldsymbol{a}\|_2^2.$$

Using this notation, we have from Eqs. (4.7) and (C.1.9) that

$$0 \leq \epsilon \leq \frac{\|\boldsymbol{a}\|_2^2}{\min(n, \|\boldsymbol{\pi}\|_1) \, \min_l \sqrt{\mathbb{E}\, d_l}}$$

and $\operatorname{Var} X_n = \Theta\left(\|\boldsymbol{a}\|_2^2\right)$, respectively. It follows that

$$\begin{aligned}
(\operatorname{Var} X_n)^{-\frac{1}{2}} \epsilon &= \mathcal{O}\left(\|\boldsymbol{a}\|_2^{-1} \frac{\|\boldsymbol{a}\|_2^2}{\min(n, \|\boldsymbol{\pi}\|_1) \, \min_l \sqrt{\mathbb{E}\, d_l}} \right) \\
&= \mathcal{O}\left(\sqrt{\frac{\|\boldsymbol{a}\|_2^2}{\min\left(n^2, \|\boldsymbol{\pi}\|_1^2\right) \, \min_l \mathbb{E}\, d_l}} \right) \\
&= \mathcal{O}\left(\sqrt{\frac{\|\boldsymbol{\pi}\|_1}{\min\left(n^2, \|\boldsymbol{\pi}\|_1^2\right) \min_l \pi_l}} \right) \quad \text{(Assumption 1)} \\
&= o\left(\sqrt{\frac{\|\boldsymbol{\pi}\|_1}{\min\left(n^{3/2}, n^{-1/2} \|\boldsymbol{\pi}\|_1^2\right)}} \right) \quad \text{(Assumption 2)} \\
&= o(1). \quad \text{(Assumptions 2 and 3)}
\end{aligned}$$

□

Lemma C.1.9. *Consider Assumptions 1 and 4. Then, whenever $K = o(n)$ it holds that*

$$\sum_{j=1}^{n}\sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij} = \Theta\Big(\|\boldsymbol{a}\|_2^2\Big).$$

Proof. Under Assumption 4, we may write

$$\begin{aligned}
\sum_{j=1}^{n}\sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij} &= \sum_{j=1}^{n}\sum_{i<j} c_{ij}^2 \Theta(\pi_i\pi_j) \\
&= \frac{1}{2}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}^2\Theta(\pi_i\pi_j) - \sum_{i=1}^{n} c_{ii}^2\Theta\big(\pi_i^2\big)\right] \\
&= \frac{1}{2}\left[\sum_{k=1}^{K}\sum_{t=1}^{K} c_{tk}^2\Theta\Big(\|\boldsymbol{\pi}\|_1^{k,\emptyset}\|\boldsymbol{\pi}\|_1^{t,\emptyset}\Big) + \mathcal{O}\Big(\|\boldsymbol{\pi}\|_2^2\Big)\right]. \quad \text{(Lemma C.1.7: } c_{ij} = \mathcal{O}(1)\text{)}
\end{aligned} \tag{C.25}$$

Recall from Lemma C.1.7 that c_{ij} can be written as a function of $g(i)$ and $g(j)$:

$$\begin{aligned}
c_{tk} &= \delta_{t=k} + \underbrace{\frac{1}{\|\boldsymbol{\pi}\|_1}\left[\sum_{l=1}^{K}\frac{\Big(\|\boldsymbol{\pi}\|_1^{l,\emptyset}\Big)^2}{\|\boldsymbol{\pi}\|_1} - \|\boldsymbol{\pi}\|_1^{t,\emptyset} - \|\boldsymbol{\pi}\|_1^{k,\emptyset}\right]}_{B} + \mathcal{O}\left(\frac{1}{n}\right) \\
\Rightarrow c_{tk}^2 &= \delta_{t=k} + 2\delta_{t=k}B + B^2 + \mathcal{O}\left(\frac{1}{n}\right). \quad \text{(Lemma C.1.7: } c_{ij} = \mathcal{O}(1)\text{)}
\end{aligned}$$

Then, in Eq. (C.25) we substitute

$$a_k = \|\boldsymbol{\pi}\|_1^{k,\emptyset},$$

(so that $\|\boldsymbol{a}\|_1 = \|\boldsymbol{\pi}\|_1$), and we obtain

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=1}^{K} c_{tk}^2\Theta(a_k a_t) \\
&= \sum_{k=1}^{K}\sum_{t=1}^{K}\left[\delta_{k=t} + 2\delta_{k=t}B + B^2 + \mathcal{O}\left(\frac{1}{n}\right)\right]\Theta(a_k a_t) \\
&= \sum_{k=1}^{K}(1+2B)\Theta\big(a_k^2\big) + \sum_{k=1}^{K}\sum_{t=1}^{K}\left[B^2 + \mathcal{O}\left(\frac{1}{n}\right)\right]\Theta(a_k a_t).
\end{aligned} \tag{C.26}$$

We now address the two terms on the right-hand side of Eq. (C.26) separately:

$$\begin{aligned}
\sum_{k=1}^{K}(1+2B)a_k^2 &= \|\boldsymbol{a}\|_2^2 + \frac{2}{\|\boldsymbol{a}\|_1}\sum_{k=1}^{K}\left(\sum_{l=1}^{K}\frac{a_l^2}{\|\boldsymbol{a}\|_1} - 2a_k\right)a_k^2 \\
&= \|\boldsymbol{a}\|_2^2 + 2\frac{\|\boldsymbol{a}\|_2^4}{\|\boldsymbol{a}\|_1^2} - 4\frac{\|\boldsymbol{a}\|_3^3}{\|\boldsymbol{a}\|_1}.
\end{aligned} \tag{C.27}$$

$$
\begin{aligned}
&\sum_{k=1}^{K}\sum_{t=1}^{K}\left[B^2+\mathcal{O}\left(\frac{1}{n}\right)\right]a_k a_t \\
&\quad=\sum_{k=1}^{K}\sum_{t=1}^{K}\left\{\frac{1}{\|\boldsymbol{a}\|_1}\left[\sum_{l=1}^{K}\frac{(a_l)^2}{\|\boldsymbol{a}\|_1}-a_k-a_t\right]\right\}^2 a_k a_t+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right) \\
&\quad=\frac{1}{\|\boldsymbol{a}\|_1^2}\sum_{k=1}^{K}\sum_{t=1}^{K}\left[\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1}-(a_k+a_t)\right]^2 a_k a_t+\mathcal{O}\left(\frac{1}{n}\right)+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right) \\
&\quad=\frac{1}{\|\boldsymbol{a}\|_1^2}\sum_{k=1}^{K}\sum_{t=1}^{K}\left[\left(\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1}\right)^2-2\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1}(a_k+a_t)+(a_k+a_t)^2\right]a_k a_t+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right) \\
&\quad=\frac{1}{\|\boldsymbol{a}\|_1^2}\left[\|\boldsymbol{a}\|_2^4-2\|\boldsymbol{a}\|_2^4+\sum_{k=1}^{K}\sum_{t=1}^{K}\left(a_k^2+2a_k a_t+a_t^2\right)a_k a_t\right]+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right) \\
&\quad=\frac{1}{\|\boldsymbol{a}\|_1^2}\left[\|\boldsymbol{a}\|_2^4-2\|\boldsymbol{a}\|_2^4+2\|\boldsymbol{a}\|_3^3\|\boldsymbol{a}\|_1+2\|\boldsymbol{a}\|_2^4\right]+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right) \\
&\quad=\frac{1}{\|\boldsymbol{a}\|_1^2}\left[\|\boldsymbol{a}\|_2^4+2\|\boldsymbol{a}\|_3^3\|\boldsymbol{a}\|_1\right]+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}\right). \qquad \text{(C.28)}
\end{aligned}
$$

Thus, substituting Eqs. (C.27) and (C.28) into Eq. (C.26) and in turn into Eq. (C.25), we obtain

$$
\begin{aligned}
\sum_{j=1}^{n}\sum_{i<j}c_{ij}^2\operatorname{Var}A_{ij}&=\Theta\left(\|\boldsymbol{a}\|_2^2+3\frac{\|\boldsymbol{a}\|_2^4}{\|\boldsymbol{a}\|_1^2}-2\frac{\|\boldsymbol{a}\|_3^3}{\|\boldsymbol{a}\|_1}\right)+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{n}+\|\boldsymbol{\pi}\|_2^2\right) \qquad \text{(C.29)}\\
&=\|\boldsymbol{a}\|_2^2\left[\Theta\left(1+3\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1^2}-2\frac{\|\boldsymbol{a}\|_2\|\boldsymbol{a}\|_3^3}{\|\boldsymbol{a}\|_1\|\boldsymbol{a}\|_2^3}\right)+\mathcal{O}\left(\frac{\|\boldsymbol{a}\|_1^2}{\|\boldsymbol{a}\|_2^2}\left\{\frac{1}{n}+\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{a}\|_1^2}\right\}\right)\right].
\end{aligned}
$$

Since $\|\boldsymbol{a}\|_1=\|\boldsymbol{\pi}\|_1$ and $\|\boldsymbol{a}\|_1^2/\|\boldsymbol{a}\|_2^2\leq K$, it follows that

$$
\begin{aligned}
&=\|\boldsymbol{a}\|_2^2\left[\Theta\left(1+3\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1^2}-2\frac{\|\boldsymbol{a}\|_2\|\boldsymbol{a}\|_3^3}{\|\boldsymbol{a}\|_1\|\boldsymbol{a}\|_2^3}\right)+\mathcal{O}\left(K\left\{\frac{1}{n}+\frac{\|\boldsymbol{\pi}\|_2^2}{\|\boldsymbol{\pi}\|_1^2}\right\}\right)\right] \\
&=\|\boldsymbol{a}\|_2^2\left[\Theta\left(1+3\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1^2}-2\frac{\|\boldsymbol{a}\|_2\|\boldsymbol{a}\|_3^3}{\|\boldsymbol{a}\|_1\|\boldsymbol{a}\|_2^3}\right)+\mathcal{O}\left(\frac{K}{n}\right)\right] \quad \text{(Assumption 1).} \qquad \text{(C.30)}
\end{aligned}
$$

Furthermore, we obtain

$$
\begin{aligned}
&\geq\|\boldsymbol{a}\|_2^2\left[\Theta\left(1+3\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1^2}-2\frac{\|\boldsymbol{a}\|_2}{\|\boldsymbol{a}\|_1}\right)+\mathcal{O}\left(\frac{K}{n}\right)\right] \\
&=\|\boldsymbol{a}\|_2^2\left[\Theta\left(\left[\sqrt{3}\frac{\|\boldsymbol{a}\|_2}{\|\boldsymbol{a}\|_1}-\frac{1}{\sqrt{3}}\right]^2+\frac{2}{3}\right)+\mathcal{O}\left(\frac{K}{n}\right)\right] \\
&=\Theta\left(\|\boldsymbol{a}\|_2^2\right). \quad (K=o(n)) \qquad \text{(C.31)}
\end{aligned}
$$

From Eq. (C.30), it follows that

$$
\sum_{j=1}^{n}\sum_{i<j}c_{ij}^2\operatorname{Var}A_{ij}\leq\|\boldsymbol{a}\|_2^2\left[\Theta\left(1+3\frac{\|\boldsymbol{a}\|_2^2}{\|\boldsymbol{a}\|_1^2}\right)+\mathcal{O}\left(\frac{K}{n}\right)\right]. \qquad \text{(C.32)}
$$

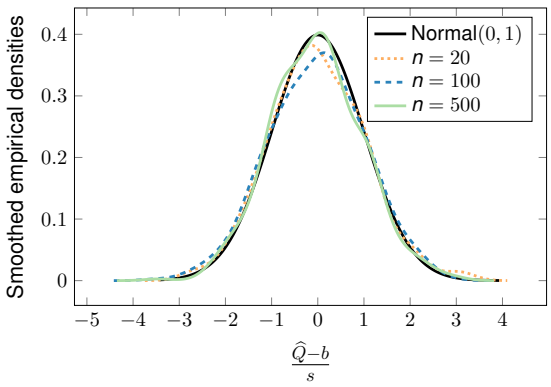
Thus, since $\|\boldsymbol{a}\|_2^2 \leq \|\boldsymbol{a}\|_1^2$, we conclude from Eqs. (C.31) and (C.32) that whenever $K = o(n)$, we obtain the required result of this lemma:

$$\sum_{j=1}^{n} \sum_{i<j} c_{ij}^2 \operatorname{Var} A_{ij} = \Theta\Big(\|\boldsymbol{a}\|_2^2\Big).$$

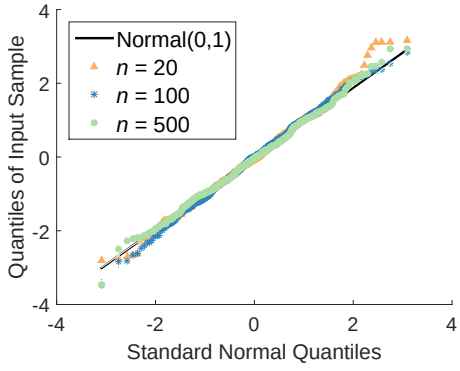
□

C.2 Simulations illustrating theorems in Chapter 4

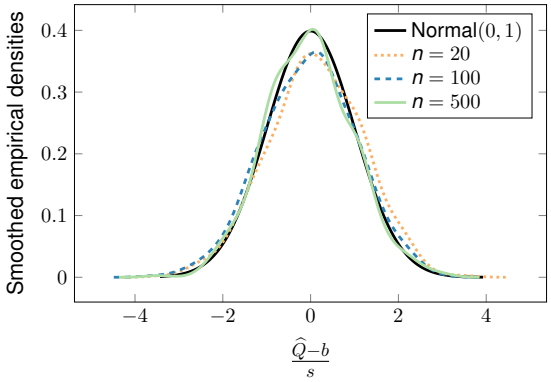
C.2.1 Simulations illustrating Theorem 4.2.2 for simple networks



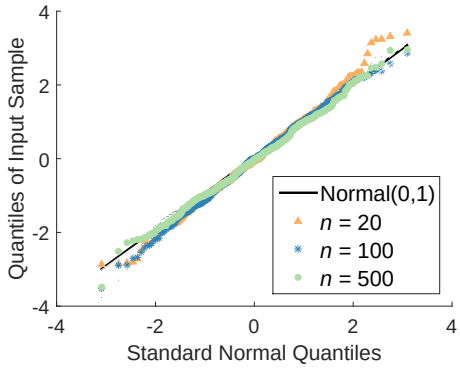
(a) Smoothed empirical densities



(b) Q-Q plot

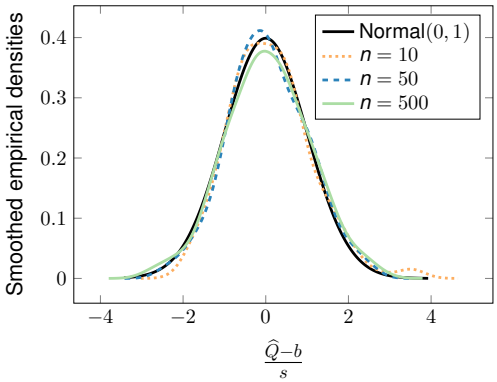


(c) Plug-in: Smoothed empirical densities

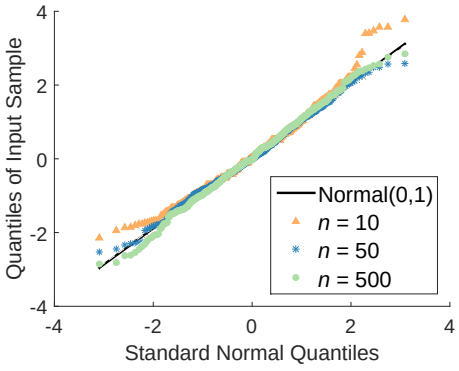


(d) Plug-in: Q-Q plot

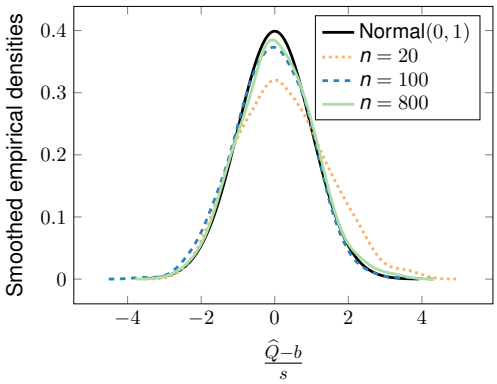
Figure C.1: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$ for simple networks; simulated from power law networks with $\mathbb{E}\, A_{ij} = (ij)^{-0.2}$.



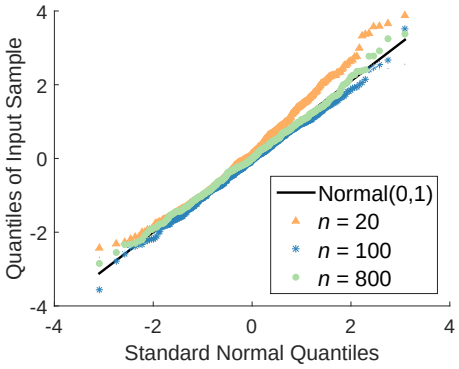
(a) Smoothed empirical densities



(b) Q-Q plot

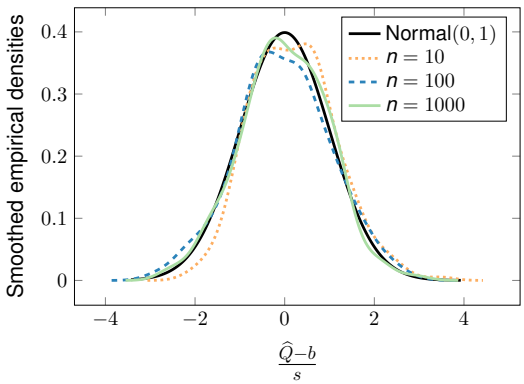


(c) Plug-in: Smoothed empirical densities

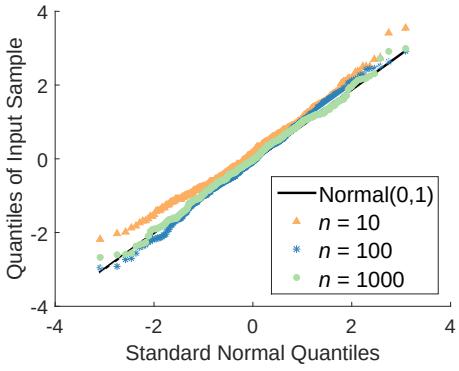


(d) Plug-in: Q-Q plot

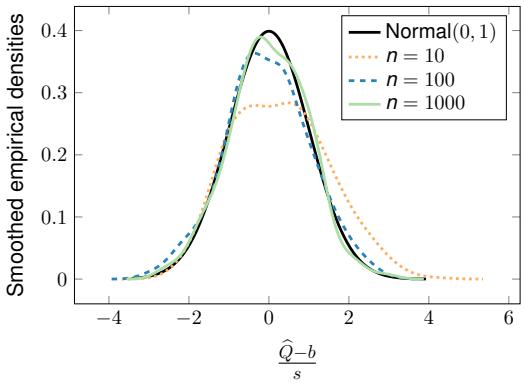
Figure C.2: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$; simulated from power law networks with $\mathbb{E}\, A_{ij} = 0.81 \cdot (ij)^{-0.7}$.



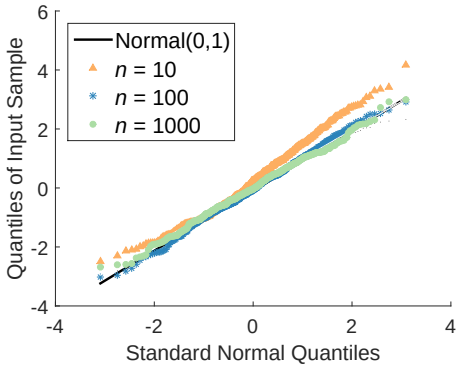
(a) Smoothed empirical densities



(b) Q-Q plot



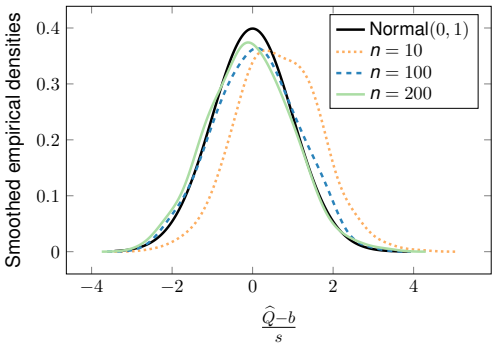
(c) Plug-in: Smoothed empirical densities



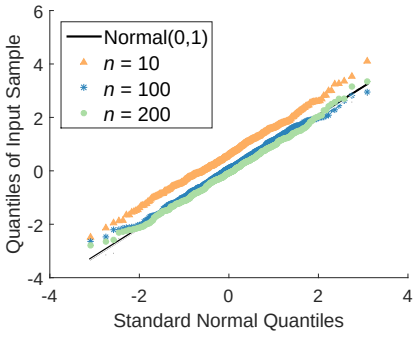
(d) Plug-in: Q-Q plot

Figure C.3: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$; simulated from power law networks with $\mathbb{E}\, A_{ij} = 0.16$.

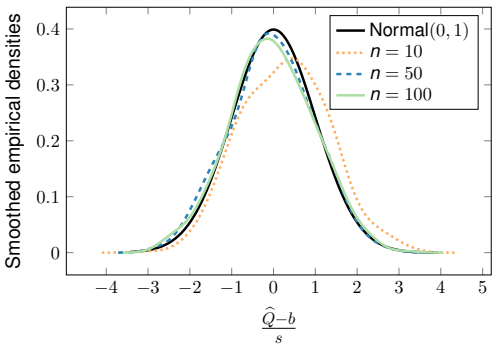
C.2.2 Simulations illustrating Theorem 4.2.2 for multi-edge networks



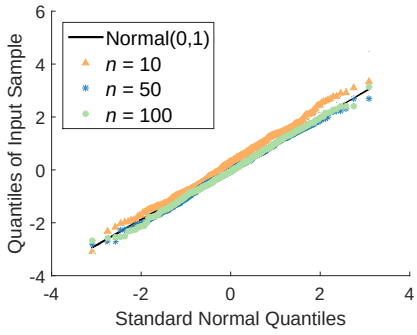
(a) Binomial: Smoothed empirical densities



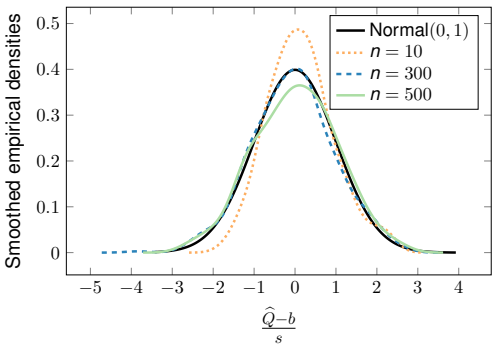
(b) Binomial: Q-Q plot



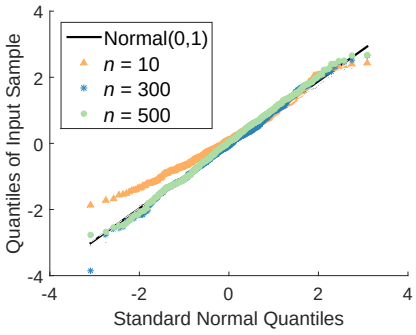
(c) Poisson: Smoothed empirical densities



(d) Poisson: Q-Q plot



(e) Negative Binomial: Smoothed empirical densities



(f) Negative Binomial: Q-Q plot

Figure C.4: Illustration of Theorem 4.2.2: The large-sample behavior of modularity $\widehat{Q}$ for multi-edge networks standardized by the plug-in estimators for b and s; simulated from power law networks with $\mathbb{E}\, A_{ij} = 11.56 \cdot (ij)^{-0.2}$.

Appendix D

Supporting material for Chapter 5

D.1 Likelihood functions for model comparison

In section 5.3, we fit and compare four models to the email interaction network: Poisson, Zero-inflated Poisson, Negative Binomial and Zero-inflated Negative Binomial. For the comparison, we will additionally need the 1-parameter model $\mathrm{Poisson}(\lambda)$ and the saturated Poisson model. We now describe each of these models in turn.

$\underline{A_{ij} \sim \mathrm{Poisson}(\lambda)}$:

We model the edges A_{ij} as independent and identical distributed Poisson random variables with the expectation $\mathbb{E}\, A_{ij} = \lambda, \forall i, j$. The corresponding log-likelihood is defined as

$$\log f(\boldsymbol{A}|\boldsymbol{\pi}) = \sum_{j=1}^{n} \sum_{i<j} [A_{ij} \log \lambda - \log(A_{ij}!) - \lambda].$$

$\underline{A_{ij} \sim \mathrm{Poisson}(\pi_i \pi_j)}$:

We model the edges A_{ij} as independent Poisson random variables with the expectation following the degree-based model of Definition 2 ($\mathbb{E}\, A_{ij} = \pi_i \pi_j, \forall i, j$). The corresponding log-likelihood is defined as

$$\log f(\boldsymbol{A}|\boldsymbol{\pi}) = \sum_{j=1}^{n} \sum_{i<j} [A_{ij} \log(\pi_i \pi_j) - \log(A_{ij}!) - \pi_i \pi_j].$$

$\underline{A_{ij} \sim \text{Zero-inflated } \mathrm{Poisson}(\pi_i \pi_j, p)}$:

We model the edges as independent zero-inflated Poisson random variables with the expectation following the degree-based model of Definition 2; i.e.,

$$A_{ij} = Z Y_{ij}, \quad Z \sim \mathrm{Bernoulli}(1-p), \quad Y_{ij} \sim \mathrm{Poisson}(\lambda'_{ij})$$

with Z and Y_{ij} being independent for all $i < j$, and $\lambda'_{ij} = \pi_i \pi_j / (1-p)$ such that $\mathbb{E}\, A_{ij} =$

$(1-p)\lambda'_{ij} = \pi_i\pi_j$. The corresponding log-likelihood is defined as

$$\log f(\boldsymbol{A}|p,\boldsymbol{\pi}) = \sum_{A_{ij}=0} \log\Big(p + (1-p)\exp\Big(-\frac{\pi_i\pi_j}{1-p}\Big)\Big) + \sum_{A_{ij}>0} \log\left((1-p)\exp\Big(-\frac{\pi_i\pi_j}{1-p}\Big)\frac{\Big(\frac{\pi_i\pi_j}{1-p}\Big)^{A_{ij}}}{A_{ij}!}\right).$$

Saturated Poisson model:

We model the edges A_{ij} as independent Poisson random variables with the expectation equal to the observed edge value ($\mathbb{E}\, A_{ij} = A_{ij}, \forall i,j$). The corresponding log-likelihood is defined as

$$\log f(\boldsymbol{A}|\boldsymbol{\pi}) = \sum_{j=1}^{n}\sum_{i<j}[A_{ij}\log(A_{ij}) - \log(A_{ij}!) - A_{ij}].$$

$A_{ij} \sim \text{NegativeBinomial}(\pi_i\pi_j, r)$:

We model the edges A_{ij} as independent Negative Binomial random variables with the expectation following the degree-based model of Definition 2 and we denote by r the shape parameter. This can be interpreted as Poisson distributed edges with mean W where we regard the mean itself as gamma distributed with $\mathbb{E}\, W = \pi_i\pi_j$ and shape parameter r. Under this parametrization we obtain $\mathbb{E}\, A_{ij} = \pi_i\pi_j$. The corresponding log-likelihood with Γ being the gamma function is defined as

$$\log f(\boldsymbol{A}|p,r,\boldsymbol{\pi}) = \sum_{A_{ij}=0} r\big[\log r - \log\big(\mu'_{ij}+r\big)\big] + \sum_{A_{ij}>0}\Bigg(\log\Gamma(A_{ij}+r) - \log\Gamma(A_{ij}-1) - \log(\Gamma(r)) + r\big[\log r - \log(\mu'_{ij}+r)\big] + A_{ij}\big[\log\mu'_{ij} - \log(\mu'_{ij}+r)\big]\Bigg).$$

$A_{ij} \sim \text{Zero-inflated NegativeBinomial}(\pi_i\pi_j, r, p)$

Furthermore, we model the edges as independent zero-inflated Negative Binomial random variables with the expectation following the degree-based model of Definition 2; i.e.,

$$A_{ij} = ZY_{ij}, \quad Z \sim \text{Bernoulli}(1-p), \quad Y_{ij} \sim \text{NegativeBinomial}(\mu'_{ij}, r)$$

with $\mu'_{ij} = \pi_i\pi_j/(1-p)$ such that $\mathbb{E}\, A_{ij} = (1-p)\mu'_{ij} = \pi_i\pi_j$. The corresponding log-likelihood

is defined as

$$\begin{aligned}\log f(\boldsymbol{A}|p,r,\boldsymbol{\pi}) = &\sum_{A_{ij}=0} \log\left[p + (1-p)\left(\frac{r}{\mu'_{ij}+r}\right)^r\right] \\ &+ \sum_{A_{ij}>0}\Bigg(\log(1-p) + \log\Gamma(A_{ij}+r) - \log\Gamma(A_{ij}-1) - \log(\Gamma(r)) \\ &\quad + r\big[\log r - \log(\mu'_{ij}+r)\big] + A_{ij}\big[\log\mu'_{ij} - \log(\mu'_{ij}+r)\big]\Bigg).\end{aligned}$$

Bibliography

[1] L. A. Adamic and N. Glance, "The political blogosphere and the 2004 U.S. election: divided they blog," *Proceedings of the 3rd international workshop on Link discovery - LinkKDD '05*, pp. 36–43, 2005.

[2] C. Aicher, A. Z. Jacobs, and A. Clauset, "Learning latent block structure in weighted networks," *Journal of Complex Networks*, vol. 3, pp. 221–248, 2015.

[3] E. M. Airoldi, D. M. Blei, S. E. Fienberg, and E. P. Xing, "Mixed membership stochastic blockmodels.," *Journal of Machine Learning Research*, vol. 9, pp. 1981–2014, 2008.

[4] D. J. Aldous, "Representations for partially exchangeable arrays of random variables," *Journal of Multivariate Analysis*, vol. 11, pp. 581–598, 1981.

[5] W. Ali, T. Rito, G. Reinert, F. Sun, and C. M. Deane, "Alignment-free protein interaction network comparison," *Bioinformatics*, vol. 30, pp. 430–437, 2014.

[6] W. Ali, A. E. Wegner, R. E. Gaunt, C. M. Deane, and G. Reinert, "Comparison of large networks with sub-sampling strategies," *Scientific Reports*, vol. 6, p. 28 955, 2016.

[7] A. A. Amini, A. Chen, P. J. Bickel, and E. Levina, "Pseudo-likelihood methods for community detection in large sparse networks," *Annals of Statistics*, vol. 41, pp. 2097–2122, 2013.

[8] E. Arias-Castro and N. Verzelen, "Community detection in dense random networks," *Annals of Statistics*, vol. 42, pp. 940–969, 2014.

[9] A. Athreya, C. E. Priebe, M. Tang, V. Lyzinski, D. J. Marchette, and D. L. Sussman, "A limit theorem for scaled eigenvectors of random dot product graphs," *Sankhya A*, vol. 78, pp. 1–18, 2016.

[10] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[11] D. S. Bassett, M. Yang, N. F. Wymbs, and S. T. Grafton, "Learning-induced autonomy of sensorimotor systems," *Nature Neuroscience*, vol. 18, pp. 744–751, 2015.

[12] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, pp. 192–236, 1974.

[13] S. Bhattacharyya and P. J. Bickel, "Subsampling bootstrap of count features of networks," *Annals of Statistics*, vol. 43, pp. 2384–2411, 2015.

[14] P. J. Bickel and A. Chen, "A nonparametric view of network models and Newman-Girvan and other modularities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, pp. 21 068–21 073, 2009.

[15] P. J. Bickel, A. Chen, and E. Levina, "The method of moments and degree distributions for network models," *Annals of Statistics*, vol. 39, pp. 2280–2301, 2011.

[16] P. J. Bickel and P. Sarkar, "Hypothesis testing for automated community detection in networks," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 78, pp. 253–273, 2016.

[17] P. Billingsley, *Probability and Measure*. New York: John Wiley & Sons, 1995.

[18] B. B. Biswal, M. Mennes, X. N. Zuo, *et al.*, "Toward discovery science of human brain function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 4734–4739, 2010.

[19] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. P10008, pp. 1–12, 2008.

[20] B. Bollobás, S. Janson, and O. Riordan, "The phase transition in inhomogeneous random graphs," *Random Structures Algorithms*, vol. 31, pp. 3–122, 2007.

[21] B. Bollobás and O. Riordan, "Metrics for sparse graphs.," *Surveys in combinatorics 2009, London Math Soc Lecture Note Series*, S. Huczynska, J. D. Mitchell, and C. M. Roney-Dougal, Eds., Cambridge: Cambridge University Press, 2009, pp. 211–287.

[22] ——, "Sparse graphs: metrics and random models," *Random Structures Algorithms*, vol. 39, pp. 1–38, 2011.

[23] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *The Journal of Mathematical Sociology*, vol. 2, pp. 113–120, 1972.

[24] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On finding graph clusterings with maximum modularity," *Graph-Theoretic Concepts in Computer Science*, A. Brandstädt, D. Kratsch, and H. Müller, Eds., vol. 4769, Berlin: Springer-Verlag, 2007, pp. 121–132.

[25] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. New York: Springer-Verlag, 1991.

[26] A. C. Cameron and P. K. Trivedi, "Econometric models based on count data: comparisons and applications of some estimators and tests," *Journal of Applied Econometrics*, vol. 1, pp. 29–53, 1986.

[27] S. Chatterjee and P. Diaconis, "Estimating and understanding exponential random graph models," *Annals of Statistics*, vol. 41, pp. 2428–2461, 2013.

[28] D. Chen, L. Lü, M.-S. Shang, Y.-C. Zhang, and T. Zhou, "Identifying influential nodes in complex networks," *Physica A: Statistical Mechanics and its Applications*, vol. 391, pp. 1777–1787, 2012.

[29] D. S. Choi, P. J. Wolfe, and E. M. Airoldi, "Stochastic blockmodels with a growing number of classes," *Biometrika*, vol. 99, pp. 273–284, 2012.

[30] F. Chung and L. Lu, "Connected Components in Random Graphs with Given Expected Degree Sequences," *Annals of Combinatorics*, vol. 6, pp. 125–145, 2002.

[31] ——, "The average distances in random graphs with given expected degrees," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 15 879–15 882, 2002.

[32] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, pp. 661–703, 2009.

[33] W. W. Cohen, "Enron email dataset," *Http://www.cs.cmu.edu/˜enron/*, accessed on 1st September, 2016.

[34] M. Coscia, F. Giannotti, and D. Pedreschi, "Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science," *Statistical Analysis and Data Mining*, vol. 4, pp. 512–546, 2011.

[35] M. Coulson, R. E. Gaunt, and G. Reinert, "Poisson approximation of subgraph counts in stochastic block models and a graphon model," *ESAIM: Probability and Statistics*, vol. 20, pp. 131–142, 2016.

[36] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human Brain Mapping*, vol. 33, pp. 1914–1928, 2012.

[37] P. Craven and B. Wellman, "The network city," *Sociological Inquiry*, vol. 43, pp. 57–88, 1973.

[38] G. Csárdi and T. Nepusz, "The igraph software package for complex network research," *InterJournal Complex Systems*, vol. 18, p. 1695, 2006.

[39] A. DasGupta, *Asymptotic Theory of Statistics and Probability*. USA: Springer-Verlag, 2008.

[40] M. De Domenico, A. Solé-Ribalta, E. Omodei, S. Gómez, and A. Arenas, "Ranking in interconnected multilayer networks reveals versatile nodes," *Nature Communications*, vol. 6, p. 6868, 2015.

[41] P. Diaconis and S. Janson, "Graph limits and exchangeable random graphs," *Rendiconti di Matematica e delle sue Applicazioni, Serie VII*, vol. 28, pp. 33–61, 2008.

[42] F. X. Diebold and K. Yilmaz, "On the network topology of variance decompositions: measuring the connectedness of financial firms," *Journal of Econometrics*, vol. 182, pp. 119–134, 2014.

[43] W. E. Donath and A. J. Hoffman, "Lower bounds for the partitioning of graphs," *IBM Journal of Research and Development*, vol. 17, pp. 420–425, 1973.

[44] D. Donoho and M. Gavish, "Minimax risk of matrix denoising by singular value thresholding," *Annals of Statistics*, vol. 42, pp. 2413–2440, 2014.

[45] J. Duch and A. Arenas, "Community detection in complex networks using extremal optimization," *Physical Review E*, vol. 72, p. 027 104, 2005.

[46] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *Journal of Human Evolution*, vol. 22, pp. 469–493, 1992.

[47] D. Durante and D. B. Dunson, "Nonparametric Bayes dynamic modeling of relational data," *Biometrika*, vol. 101, pp. 883–898, 2014.

[48] R. Durrett, *Random Graph Dynamics*. Cambridge, UK: Cambridge University Press, 2007.

[49] P. Erdös and A. Rényi, "On random graphs," *Publicationes Mathematicae*, vol. 6, pp. 290–297, 1959.

[50] I. Fellows and M. S. Handcock, "Exponential-family random network models," *Unpublished manuscript, arXiv:1208.0121*, 2012.

[51] M. Fiedler, "Algebraic connectivity of graphs," *Czechoslovak Mathematical Journal*, vol. 23, pp. 298–305, 1973.

[52] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 36–41, 2007.

[53] B. K. Fosdick and P. D. Hoff, "Testing and modeling dependencies between a network and nodal attributes," *Journal of American Statistical Association*, vol. 110, pp. 1047–1056, 2015.

[54] O. Frank and D. Strauss, "Markov graphs," *Journal of American Statistical Association*, vol. 81, pp. 832–842, 1986.

[55] B. Franke, J.-F. Plante, R. Roscher, A. Lee, C. Smyth, A. Hatefi, F. Chen, E. Gil, A. Schwing, A. Selvitella, M. M. Hoffman, R. Grosse, D. Hendricks, and N. Reid, "Statistical inference, learning and models in big data," *International Statistical Review*, doi:10.1111/insr.12176, 2016.

[56] B. Franke and P. J. Wolfe, "Network modularity in the presence of covariates," *Unpublished manuscript, arXiv:1603.01214*, 2016.

[57] L. C. Freeman, "A set of measures of centrality based on betweenness," *Sociometry*, vol. 40, pp. 35–41, 1977.

[58] E. N. Gilbert, "Random graphs," *Annals of Mathematical Statistics*, vol. 30, pp. 1141–1144, 1959.

[59] P. M. Gleiser and L. Danon, "Community structure in jazz," *Advances in Complex Systems*, vol. 6, pp. 565–573, 2003.

[60] D. Godwin, R. L. Barry, and R. Marois, "Breakdown of the brain's functional network modularity with awareness," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 3799–3804, 2015.

[61] B. H. Good, Y.-A. de Montjoye, and A. Clauset, "Performance of modularity maximization in practical contexts," *Physical Review E*, vol. 81, p. 046 106, 2010.

[62] M. S. Handcock, "Assessing degeneracy in statistical models of social networks," *Working Paper No. 39, Center for Statistics and the Social Sciences, University of Washington, Seattle*, 2003.

[63] M. S. Handcock, A. E. Raftery, and J. M. Tantrum, "Model-based clustering for social networks," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, vol. 170, pp. 301–354, 2007.

[64] S. Hanneke, W. Fu, and E. P. Xing, "Discrete temporal models of social networks," *Electronic Journal of Statistics*, vol. 4, pp. 585–605, 2010.

[65] T. Harford, "Big data: are we making a big mistake?" *Financial Times*, 24th March, 2014.

[66] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *Journal of the American Statistical Association*, vol. 100, pp. 286–295, 2005.

[67] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, vol. 97, pp. 1090–1098, 2002.

[68] P. W. Holland, K. B. Laskey, and S. Leinhardt, "Stochastic blockmodels: first steps," *Social Networks*, vol. 5, pp. 109–137, 1983.

[69] D. N. Hoover, "Relations on probability spaces and arrays of random variables," Institute for Advanced Study, Princeton, Tech. Rep., 1979.

[70] A. L. Hopkins, "Network pharmacology: the next paradigm in drug discovery.," *Nature Chemical Biology*, vol. 4, pp. 682–690, 2008.

[71] D. R. Hunter, S. M. Goodreau, and M. S. Handcock, "Goodness of fit of social network models," *Journal of American Statistical Association*, vol. 103, pp. 248–258, 2008.

[72] D. R. Hunter and M. S. Handcock, "Inference in curved exponential family models for networks," *Journal of Computational and Graphical Statistics*, vol. 15, pp. 565–583, 2006.

[73] N. L. Johnson, A. W. Kemp, and S. Kotz, *Univariate Discrete Distributions*, 3rd ed. New Jersey: John Wiley & Sons, 2005.

[74] A. Joseph and B. Yu, "Impact of regularization on spectral clustering," *Annals of Statistics*, vol. 44, pp. 1765–1791, 2016.

[75] B. Karrer and M. E. J. Newman, "Stochastic blockmodels and community structure in networks," *Physical Review E*, vol. 83, p. 016 107, 2011.

[76] M. Kearns, A. Roth, Z. S. Wu, and G. Yaroslavtsev, "Private algorithms for the protected in social network search," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 913–918, 2016.

[77] M. Kivelä, A. Arenas, M. Barthélemy, J. P. Gleeson, Y. Moreno, and M. A. Porter, "Multilayer networks," *Journal of Complex Networks*, vol. 2, pp. 203–271, 2014.

[78] K.-K. Kleineberg, M. Boguñá, M. Ángeles Serrano, and F. Papadopoulos, "Hidden geometric correlations in real multiplex networks," *Nature Physics*, DOI: 10.1038/NPHYS3812, 2016.

[79] E. D. Kolaczyk, *Statistical Analysis of Network Data: Methods and Models*. New York: Springer-Verlag, 2009.

[80] P. N. Krivitsky and M. S. Handcock, "A separable model for dynamic networks," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 76, pp. 26–46, 2014.

[81] P. N. Krivitsky, M. S. Handcock, A. E. Raftery, and P. D. Hoff, "Representing degree distributions, clustering, and homophily in social networks with latent cluster random effects models," *Social Networks*, vol. 31, pp. 204–213, 2009.

[82] P. Latouche, E. Birmelé, and C. Ambroise, "Overlapping stochastic block models with application to the French political blogosphere," *Annals of Applied Statistics*, vol. 5, pp. 309–336, 2011.

[83] P. Latouche, S. Robin, and S. Ouadah, "Goodness of fit of logistic models for random graphs," *Unpublished manuscript, arXiv:1508.00286*, 2015.

[84] N. H. Lee and C. E. Priebe, "A latent process model for time series of attributed random graphs," *Statistical Inference for Stochastic Processes*, vol. 14, pp. 231–253, 2011.

[85] E. L. Lehmann, *Elements of Large-Sample Theory*. New York: Springer-Verlag, 1999.

[86] J. Lei, "A goodness-of-fit test for stochastic block models," *Annals of Statistics*, vol. 44, pp. 401–424, 2016.

[87] J. Lei and A. Rinaldo, "Consistency of spectral clustering in stochastic block models," *Annals of Statistics*, vol. 43, pp. 215–237, 2015.

[88] E. A. Leicht and M. E. J. Newman, "Community structure in directed networks," *Physical Review Letters*, vol. 100, p. 118 703, 2008.

[89] T. Li, E. Levina, and J. Zhu, "Regression with network cohesion," *Unpublished manuscript, arXiv:1602.01192*, 2016.

[90] J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy, "Random function priors for exchangeable arrays with applications to graphs and relational data," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[91] S. Lohr, "The age of big data," *The New York Times*, 11th February, 2012.

[92] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[93] E. Maggioni, M. G. Tana, F. Arrigoni, C. Zucca, and B. A. M, "Constructing fMRI connectivity networks: a whole brain functional parcellation method for node definition," *Journal of Neuroscience Methods*, vol. 228, pp. 86–99, 2014.

[94] D. Marbach, J. C. Costello, R. Küffner, *et al.*, "Wisdom of crowds for robust gene network inference," *Nature Methods*, vol. 9, pp. 796–804, 2012.

[95] M. Martino, P. Magioncalda, Z. Huang, B. Conio, N. Piaggio, N. W. Duncan, G. Rocchi, A. Escelsior, V. Marozzi, A. Wolff, M. Inglese, M. Amore, and G. Northoff, "Contrasting variability patterns in the default mode and sensorimotor networks balance in bipolar depression and mania.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 4824–4829, 2016.

[96] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, 2nd ed. London: Chapman and Hall, 1983.

[97] C. McDiarmid and F. Skerman, "Modularity in random regular graphs and lattices," *Electronic Notes in Discrete Mathematics*, vol. 43, pp. 431–437, 2013.

[98] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual Review of Sociology*, vol. 27, pp. 415–444, 2001.

[99] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 61–67, 1967.

[100] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, pp. 824–827, 2002.

[101] J. L. Moreno, "Sociometry in relation to other social sciences," *American Sociological Association*, vol. 1, pp. 206–219, 1937.

[102] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela, "Community structure in time-dependent, multiscale, and multiplex networks," *Science*, vol. 328, pp. 876–879, 2010.

[103] National Cancer Institute, "Genomic data commons," *Https://gdc.nci.nih.gov/*, accessed on 25th August, 2016.

[104] M. E. J. Newman, "The structure of scientific collaboration networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, pp. 404–409, 2001.

[105] ——, "Analysis of weighted networks," *Physical Review E*, vol. 70, p. 056 131, 2004.

[106] ——, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, p. 066 133, 2004.

[107] ——, "Finding community structure in networks using the eigenvectors of matrices," *Physical Review E*, vol. 74, p. 036 104, 2006.

[108] ——, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 8577–8582, 2006.

[109] ——, "Community detection in networks: modularity optimization and maximum likelihood are equivalent," *Unpublished manuscript, arXiv:1606.02319*, 2016.

[110] M. E. J. Newman and A. Clauset, "Structure and inference in annotated networks," *Unpublished manuscript, arXiv:1507.04001*, 2015.

[111] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, p. 026 113, 2004.

[112] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, p. 036 122, 2003.

[113] M. E. J. Newman and G. Reinert, "Estimating the number of communities in a network," *Unpublished manuscript, arXiv:1605.02753*, 2016.

[114] M. E. J. Newman, D. J. Watts, and S. H. Strogatz, "Random graph models of social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 2566–2572, 2002.

[115] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American Statistical Association*, vol. 96, pp. 1077–1087, 2001.

[116] S. C. Olhede and P. J. Wolfe, "Degree-based network models," *Unpublished manuscript, arXiv:1211.6537*, 2012.

[117] ——, "Network histograms and universality of blockmodel approximation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 14 722–14 727, 2014.

[118] E. L. Paluck, H. Shepherd, and P. M. Aronow, "Changing climates of conflict: a social network experiment in 56 schools," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 566–571, 2016.

[119] P. O. Perry and P. J. Wolfe, "Null models for network data," *Unpublished manuscript, arXiv:1201.5871*, 2012.

[120] ——, "Point process modelling for directed interaction networks," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 75, pp. 821–849, 2013.

[121] M. Ramot, S. Grossman, D. Friedman, and R. Malach, "Covert neurofeedback without awareness shapes cortical network spontaneous connectivity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, E2413–2420, 2016.

[122] M. Rask-Andersen, M. S. Almén, and H. B. Schiöth, "Trends in the exploitation of novel drug targets," *Nature Reviews: Drug discovery*, vol. 10, pp. 579–590, 2011.

[123] J. Reichardt and S. Bornholdt, "Statistical mechanics of community detection," *Physical Review E*, vol. 74, p. 016 110, 2006.

[124] M. D. Resnick, P. S. Bearman, R. W. Blum, K. E. Bauman, K. M. Harris, J. Jones, J. Tabor, T. Beuhring, R. E. Sieving, M. Shew, M. Ireland, L. H. Bearinger, and J. R. Udry, "Protecting adolescents from harm: findings from the National Longitudinal Study on Adolescent Health," *Journal of the American Medical Association*, vol. 278, pp. 823–832, 1997.

[125] L. F. Robinson and C. E. Priebe, "Detecting time-dependent structure in network data via a new class of latent process models," *Unpublished manuscript, arXiv:1212.35871*, 2012.

[126] K. Rohe, S. Chatterjee, and B. Yu, "Spectral clustering and the high-dimensional stochastic blockmodel," *Annals of Statistics*, vol. 39, pp. 1878–1915, 2011.

[127] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: uses and interpretations," *NeuroImage*, vol. 52, pp. 1059–1069, 2010.

[128] G. Sabidussi, "The centrality index of a graph," *Psychometrika*, vol. 31, pp. 581–603, 1966.

[129] P. Sarkar and P. J. Bickel, "Role of normalization in spectral clustering for stochastic blockmodels," *Annals of Statistics*, vol. 43, pp. 962–990, 2015.

[130] F. Scharnowski, C. Hutton, O. Josephs, N. Weiskopf, and G. Rees, "Improving visual perception through neurofeedback.," *Journal of Neuroscience*, vol. 32, pp. 17 830–17 841, 2012.

[131] D. K. Sewell and Y. Chen, "Latent space models for dynamic networks," *Journal of the American Statistical Association*, vol. 110, pp. 1646–1657, 2015.

[132] K. Shibata, T. Watanabe, Y. Sasaki, and M. Kawato, "Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation," *Science*, vol. 334, pp. 1413–1415, 2011.

[133] G. Simons and Y.-C. Yao, "Asymptotics when the number of parameters tends to infinity in the Bradley-Terry model for paired comparisons," *Annals of Statistics*, vol. 27, pp. 1041–1060, 1999.

[134] T. A. B. Snijders, "Stochastic actor-oriented models for network change," *The Journal of Mathematical Sociology*, vol. 21, pp. 149–172, 1996.

[135] T. A. B. Snijders and K. Nowicki, "Estimation and prediction for stochastic blockmodels for graphs with latent block structure," *Journal of Classification*, vol. 14, pp. 75–100, 1997.

[136] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specifications for exponential random graph models," *Sociological Methodology*, vol. 36, pp. 99–153, 2006.

[137] D. L. Sussman, M. Tang, D. E. Fishkind, and C. E. Priebe, "A consistent adjacency embedding for stochastic blockmodel graphs," *Journal of the American Statistical Association*, vol. 107, pp. 1119–1128, 2012.

[138] D. L. Sussman, M. Tang, and C. E. Priebe, "Consistent latent position estimation and vertex classification for random dot product graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 48–57, 2014.

[139] M. Szell, R. Lambiotte, and S. Thurner, "Multirelational organization of large-scale social networks in an online world," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, pp. 13 636–13 641, 2010.

[140] M. Tang, A. Athreya, D. L. Sussman, V. Lyzinski, and C. E. Priebe, "A semiparametric two-sample hypothesis testing problem for random graphs," *Journal of Computational and Graphical Statistics*, DOI: 10.1080/10618600.2016.1193505, 2016.

[141] A. Vinayagam, T. E. Gibson, H.-J. Lee, B. Yilmazel, C. Roesel, Y. Hu, Y. Kwon, A. Sharma, Y.-Y. Liu, N. Perrimon, and A.-L. Barabási, "Controllability analysis of the directed human protein interaction network identifies disease genes and drug targets," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 4976–4981, 2016.

[142] A. Volfovsky and P. D. Hoff, "Testing for nodal dependence in relational data matrices," *Journal of American Statistical Association*, vol. 110, pp. 1037–1046, 2015.

[143] Y. X. R. Wang and P. J. Bickel, "Likelihood-based model selection for stochastic block models," *Unpublished manuscript, arXiv:1502.02069*, 2015.

[144] S. Wasserman and P. Pattison, "Logit models and logistic regressions for social networks: I. an introduction to Markov graphs and p*," *Psychometrika*, vol. 61, pp. 401–425, 1996.

[145] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks.," *Nature*, vol. 393, pp. 440–442, 1998.

[146] B. Wellman, "Is Dunbar's number up?" *British Journal of Psychology*, vol. 103, pp. 174–176, 2012.

[147] A. H. Westveld and P. D. Hoff, "A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict," *Annals of Applied Statistics*, vol. 5, pp. 843–872, 2011.

[148] S. Wu, A. Joseph, A. S. Hammonds, S. E. Celniker, B. Yu, and E. Frise, "Stability-driven nonnegative matrix factorization to interpret spatial gene expression and build local gene networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 113, pp. 4290–4295, 2016.

[149] E. P. Xing, W. Fu, and L. Song, "A state-space mixed membership blockmodel for dynamic network tomography," *Annals of Applied Statistics*, vol. 4, pp. 535–566, 2010.

[150] Ö. N. Yaveroglu, T. Milenković, and N. Pržulj, "Proper evaluation of alignment-free network comparison methods.," *Bioinformatics*, vol. 31, pp. 2697–2704, 2015.

[151] S. Young and E. Scheinerman, "Random dot product graph models for social networks," *Algorithms and models for the web-graph*, A. Bonato and F. R. K. Chung, Eds., vol. 4863, Berlin: Springer-Verlag, 2007, pp. 138–149.

[152] N. Zerubavel, P. S. Bearman, J. Weber, and K. N. Ochsner, "Neural mechanisms tracking popularity in real-world social networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, pp. 15 072–15 077, 2015.

[153] Y. Zhang, E. Levina, and J. Zhu, "Community detection in networks with node features," *Unpublished manuscript, arXiv:1509.01173*, 2015.

[154] Y. Zhao, E. Levina, and J. Zhu, "Consistency of community detection in networks under degree-corrected stochastic block models," *Annals of Statistics*, vol. 40, pp. 2266–2292, 2012.

[155] K. Zhou, H. K. Pedersen, A. Y. Dawed, and E. R. Pearson, "Pharmacogenomics in diabetes: insights into drug action and drug discovery," *Nature Reviews: Endocrinology*, vol. 12, pp. 337–346, 2016.